Preface

Lin Chao
Editor
Intel Technology Journal

This Q4'98 issue of the Intel Technology Journal (ITJ) describes Intel's manufacturing processes and strategies. Semiconductor manufacturing is characterized by very complex process flows made up of many process steps, all built to very close tolerances. Furthermore, there are complex interactions in these process flows. The papers in this issue describe how Intel develops components technology and manufacturing capability to deliver high-performance, cost-effective, quality products.

In 1965, Gordon Moore, co-founder of Intel, was preparing a speech and made a memorable observation. He observed that device complexity doubles about every 18 months. This observation is now known as Moore's Law. While originally intended as a rule of thumb, it has become the guiding principle for the industry spearheading the delivery of ever-more-powerful semiconductor chips at proportionate decreased costs. Intel has expended enormous resources to meet the predictions of Moore's Law through factory modeling, knowledge management, operational and simulation modeling, capacity supply line management using Goldratt's Theory of Constraints, and defect yield monitoring. These techniques are described in detail in this issue.

Among these techniques is Intel's proven Copy EXACTLY! methodology for factory ramp and high-volume manufacturing, which is described in detail in this issue. Copy EXACTLY! enables Intel to bring factories on-line quickly with high-volume practices already in place; hence, decreasing time to market and increasing yields. Copy EXACTLY! solves the problem of getting production facilities up to speed quickly by "copying" everything--process flows, equipment set, suppliers, plumbing, manufacturing clean room, and even training methodologies--from the development plant to the volume-manufacturing plant.

As we near the new millennium, the semiconductor industry is getting ready for the transition from 200mm to 300mm wafer size. This is a major milestone for the industry, and teams are in place at Intel to implement the transition. An interesting retrospective look at how the 300mm standard was selected is described in another paper in this issue.

Intel's Manufacturing—a Competitive Edge

By Mike Splinter ,
Corporate Vice President and General Manager, Technology and Manufacturing Group
Intel Corp.

Intel has a long history of technology innovations that have led the industry in establishing new capabilities in semiconductor technologies thereby enabling advances in computing. One of our successes in helping drive technology has been our ability to incorporate innovations in our manufacturing sciences. This is a critical strategy as technology alone will not ensure that products can be built at the right volumes with the right costs and then delivered to our customers at the right time.

During the 80's, we were driven to find methods to improve our manufacturing competence. Our competitors were achieving higher yields, transferring and ramping new technologies faster than us, and, overall, running their factories more efficiently. To remain competitive, we made the basics of manufacturing technology a key piece of our agenda.

One of the main issues we faced was the phenomenon of something called the Intel U. The Intel U was a predictable drop in factory performance every time a new technology or product was developed and transferred to manufacturing. This phenomenon, when plotted as a function of yield versus time, always showed a U-shaped curve. There were many approaches taken over the years to fix this problem but until the concept of Copy Exactly! was developed, we had limited success in eliminating the Intel U. Today, this phenomenon is non-existent, and Intel U stands for Intel University, not a yield problem.

There have been many innovations in manufacturing that have addressed, among others, defect improvements, modeling, capacity management, and improving the speed of our supply line. Today, Intel is faced with the challenges of developing new concepts and methods in manufacturing in order to address the changes we have seen in the marketplace. The world is moving at Internet speed fostering the concepts of build to order, E-Commerce, product segmentation, and so on. As we move into the future, we will need to focus not only on TECHNOLOGY, but also on COST, SPEED, AND FLEXIBILITY in order for manufacturing to continue to provide Intel with a key competitive advantage over the next decade.

21st Century Semiconductor Manufacturing Capabilities

Eugene S. Meieran, Intel Fellow
Technology Manufacturing Engineering, Intel Corp.

Index words: manufacturing, operational cost, NGM program

Abstract

Semiconductor device manufacturers face many difficult challenges as we enter the 21st century. Some are direct consequences of adherence to Gordon Moore's Law, which states that device complexity doubles about every 18 months. Feature size reduction, increased wafer diameter, increased chip size, ultra-clean processing, and defect reduction among others are manifestations that have a direct bearing on the cost and quality of products, factory flexibility in responding to changing technology or business conditions, and on the timelines of product delivery to the ultimate customer.

Regardless of these tremendously complex problems, the industry is focused on meeting the predictions of Moore's Law, for which enormous resources are expended.

One of the great challenges ascribed to Moore's Law, that facility costs increase on a semi-log scale, is now known as Moore's Second Law. However, unlike his First Law, the industry would prefer to depart from Second Law predictions to avoid hugely expensive ($20 Billion) future fabs and attendant high chip costs. Logistics control, inventory management, better facility design, supplier management programs, etc. are all responses to Second Law predictions, to which many resources have been devoted.

Other pressures on factory management are emerging. In addition to cost considerations, reduction in feature size and increasingly complex devices, the Massachusetts Institute of Technology/Leaders For Manufacturing-led program, "Next Generation Manufacturing" (NGM) identifies the following issues as significant:

- globalization of supplier, customer, and factory base
- exponential growth of information and knowledge management capabilities that enable faster and better decisions
- development of new materials and processes at atomic scale dimensions
- faster delivery of higher quality products to an increasingly demanding customer
- rising awareness of environmental and energy concerns

This paper discusses the technological responses of industry management and university faculty to the predictions of Moore's Second Law. Special attention is given to knowledge management and operational modeling and simulation technology. These processes help us better understand the benefits of various alternatives used to affect factory performance as traditional methods such as yield improvement, automation, increased wafer size, equipment reliability, etc. lose their leverage.

Introduction

Gordon Moore first proposed the law that bears his name in the late '60's: chip complexity (as defined by the number of active elements on a single semiconductor chip) will double about every device generation, usually taken as about 18 calendar months [1]. This law has now been valid for more than three decades, and it appears likely to be valid for several more device generations, as shown in Figure 1.

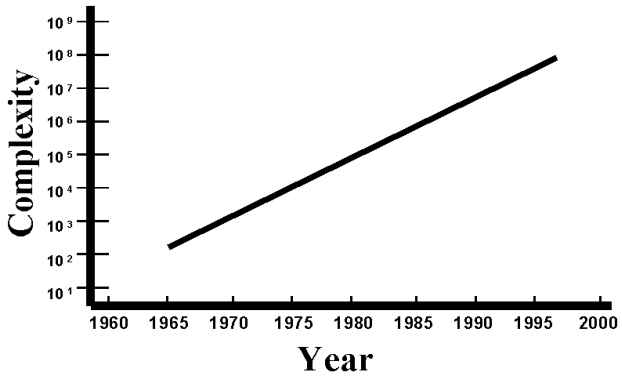

Figure 1: Moore's First Law

The compelling desire of the semiconductor industry to follow Moore's Law has affected high-volume device manufacturing, driving both the cost per bit of the devices and the overall cost of the fabrication and assembly facilities needed to build them. (Additional effects such as those on the ramp rate towards high-volume manufacturing are also experienced, but these are not discussed in this paper.)

For Moore's Law to remain valid, feature size must continue to be reduced, but since this reduction is insufficient in and of itself, chip size must continue to increase. Together, these two trends have not only maintained Moore's Law, but have accounted for the phenomenal success of our industry, since the cost per device element has now decreased by several orders of magnitude! Compared to *every* other commodity in the world, semiconductor chips are *cheap*, and continue to get cheaper (on a per element basis) every year.

The reduction in cost per active chip element is shown in Figure 2. Notice that while this cost continues to decrease, there appears to be a break in the curve: one section follows early predictions of Moore's Law, and the other departs from these predictions. This will be discussed later.

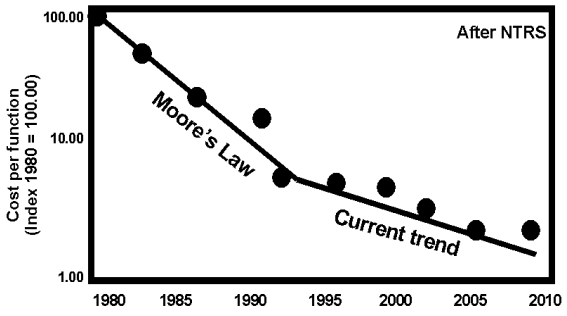

Figure 2: Cost per chip element

Many programs are associated with following Moore's Law and each has consequences for the cost per chip element, as shown in Table 1.

	1975	1997	2003
Chip complexity (index to 1)	1	10	100
Feature size reduction, μ m	2	0.25	0.08
Chip size increase, mm^2	30	150	600
Wafer diameter, mm	50	200	300
Facility automation, %	5	60	80
Die yield, % good	40	85	95
Line yield, % good	40	90	95
Assembly/test yield, %	90	99	99
Defect levels, DPM	2%	500	50

Table 1: Programs to maintain Moore's Law

Most of these programs tend to contribute to a reduction in chip element cost, but some of them, especially those dealing directly with increased chip and process complexity, tend to increase that cost. Fortunately, scaling, reduced feature size, improved yield, and increased wafer diameter more than make up for the added costs of more expensive packages and more complex processing.

Figure 3 shows the other major consequence of following Moore's Law. The reduction in cost per chip element is just offset by the increase in element density, leading to an essentially constant cost per individual chip. However, as a result, overall factory costs increase almost exponentially as we struggle to meet the ever increasing demand for more and larger high-performance chips. In order to meet cost per chip goals, cost per factory has increased to the point where high-volume factories now cost several billion dollars! So being successful in reducing chip costs brings its own share of additional problems. Building, equipping, and maintaining billion dollar factories tax even the most successful companies. This explosion of factory cost has come to be known as Moore's Second Law—one we do NOT wish to follow with such great zeal!

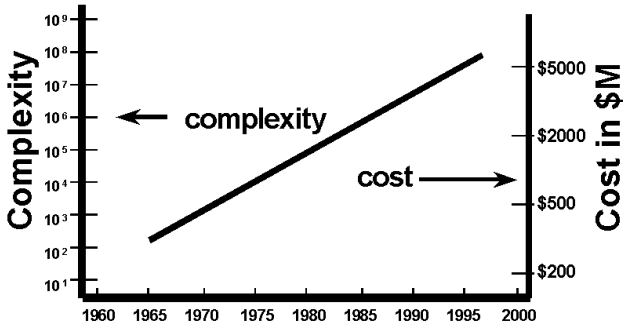

Figure 3: Moore's Second Law

Many of the same programs that have driven cost per chip element down are also responsible for the trend shown in Figure 3. In addition, some operational programs that have had little direct influence on cost per chip element have significant influence on factory cost. These additional programs are shown in Table 2.

	1975	1997	2003
Chip complexity (index to 1)	1	10	100
Feature size reduction, μ m	2	0.25	0.08
Chip size increase, mm^2	30	150	600
Wafer diameter, mm	50	200	300
Facility automation, %	5	60	80
Die yield, % good	40	85	95
Line yield, % good	40	90	95
Assembly/test yield, %	90	99	99
Operational efficiency	1	10	100
Equipment cost	1	10	50
Defect levels, DPM	2%	500	50

Table 2: Factory cost control programs

Tables 1 and 2 are combined, below, in Table 3, which shows two emerging problems with regard to both cost per chip element and factory cost containment:

1. Some goals of the programs are in conflict: lowering the cost per element actually adds to factory cost.
2. The leverage of some of the programs is diminishing: for example, we will not exceed 100% yield or 100% automation.

Hence, other means are necessary to meet cost projection goals for factories and chip elements.

	Cost per function	Factory Cost
Complexity increase	Up	Up
Feature size reduction	Down	Up
Chip size increase	Down	Up
Wafer size increase	Down	Slowing
Facility automation	Down	Slowing
Die yield	Down	Slowing
Line yield	Down	Slowing
Assembly/test yield	Down	Even
Operational efficiency	Down	Down

Table 3: Comparison of programs

The major program that does **not** suffer from topping out or from conflict is improving operational efficiency. However, before we discuss this, some additional forces acting on the manufacturing environment are discussed.

Emerging Trends

The additional forces acting on the manufacturing environment have little to do with Moore's Law. These forces are discussed in the National Science Foundation sponsored program, "Next Generation Manufacturing" conducted by the Leaders For Manufacturing program at MIT, the Agility Forum and the Technology to Enable Lean Manufacturing [2]. The major issues are listed in Table 4.

Globalization refers to the fact that for a number of reasons, industries are locating manufacturing facilities in many geographical locations, utilizing a supply of skilled workers at reasonable wages and servicing a widely dispersed customer base. As a result, suppliers of parts, materials, and equipment for these factories have had to become globalized, since operating conditions for manufacturers dictate that short time to delivery to the local customer is a matter of competitive necessity.

•	**Manufacturing globalization:**
	- Factories
	- Suppliers
	- Customers
•	**Increased global competition**
•	**Increased customer expectations**
•	**New technologies and processes**
•	**Environmentally aware manufacturing**
•	**Human factors:**
	- Training and retraining
	- Redeployment
	- Organizational structure
	- Wages and reward structure
	- Globally dispersed collaboration
•	**Pervasive information technology:**
	- Computation
	- Communication

Table 4: Emerging manufacturing needs

Due to the pervasive and timely availability of information and knowledge, global competition is more aggressive: new products are developed and brought to market quickly to globally distributed customers. Consequently, there is an erosion of what had been known as customer loyalty. Just as industrial jobs are no longer secure for life, brand-name loyalty on the part of a customer is not likely to survive; customers shop around for the most convenient or persuasive supplier.

Customers' expectations are increasing: they expect on-time delivery of high-quality customized products at prices reflective of high-volume manufacturing costs, and great service; otherwise, they will find other suppliers without hesitation. Quality is a given, not a differentiator; if one producer's product does not exhibit high quality, the customer will quickly find someone else.

Environmental concerns are also becoming more important in response to government regulations and societal concerns. Industry is recognizing that environmentally sound manufacturing is more rewarding than environmentally insensitive manufacturing.

Firms expecting to compete in the next millennium will have to play this ball game, on this playing field, with these new rules, encumbered as well by the needs and requirements listed in Table 3. These are the challenges the semiconductor industry faces as markets change, customer requirements change, and political and socio-economic forces affect how business is carried out.

Information Technology Responses

Two items listed in Table 4 were not discussed above: new technologies, materials, and processes; and greater access

to global information and knowledge. The first refers to the fact that we can now create materials and structures on an atomic scale, with properties hitherto not only unavailable, but undreamed of. New products such as micro-motors, micro-refrigerators, micro-turbines, device analysis tools, and packaging will probably generate significant business in the not-too-distant future. However, since these do not concern semiconductor chip costs at the moment, they are not discussed further here. For details on these opportunities, see the NGM report [2].

The explosion of information technology (IT) is however another story. Indeed, information technology—the pervasive generation, storage, distribution and use of information and knowledge—seems to be the technology that may help resolve ALL the dilemmas of cost and competitiveness. IT can help with the declining rate of cost per chip element and increased cost per factory, as well as those emerging manufacturing needs identified in Table 4. In the remainder of this paper, we discuss how specific elements of information technology can be used to significantly impact all these issues.

Two applications of information technology that appear to have the greatest leverage are operational modeling and simulation, and management of knowledge assets and intellectual capital. In addition, these programs also affect the third way of decreasing these costs, the ramp speed to high-volume manufacturing. (For example, Intel's Copy EXACTLY! policy is one way of managing our corporate knowledge and wisdom to increase ramp speed.) However, ramp rate improvement is not discussed further in this paper. Please refer to "The Evolution of Intel's Copy EXACTLY! Technology Transfer Method" in this issue of the *Intel Technology Journal* for a fuller description of this important program.

Operational modeling and simulation (OM&S) and management of knowledge assets and intellectual capital applications have different purposes. OM&S is used to lower the cost and speed up the process of trying alternative solutions to different operational scenarios. It can provide quicker and more accurate answers to questions such as how much equipment or how many people are needed to perform a given number of activities; how can a factory be laid out for improved efficiency; how can equipment be best located to provide high throughput and still be easily accessible for maintenance; or how equipment operation can be best scheduled to improve overall capital utilization. In order to answer these questions, different alternatives can be tried out on the computer, saving months or years of physical experimentation time, and millions or even tens of millions of dollars of experimental materials and equipment time.

In Knowledge Management (KM), ever more transient users can access vast sources of data, information and knowledge in real time to enable them to make more informed and higher quality decisions. This information is wide in scope and sufficiently deep to enable one versed in the use of such technology to make and execute decisions with unparalleled ability. Considering that the value of a corporation is more and more dependent on intellectual assets (patents, know-how, trade secrets, processing and product knowledge, best-known methods, etc.) than on capital assets (equipment, buildings, rights of way, etc.) it is not surprising that significant attention is now being paid to knowledge management.

Both OM&S and KM can be beneficially applied to the many domains of manufacturing including scheduling, using the theory of constraints tied to operational models; enterprise integration tied to enterprise models; electronic commerce; capacity planning and factory layout improvements, tied to comprehensive factory models; improved equipment utilization and performance, tied to equipment and material handling models. All these domains can benefit significantly from OM&S and KM. Using these methods, we can now start to overcome some of the limitations we face as yields approach 100%, as factory automation approaches an economical limit, and as increased wafer diameter and increased package complexity continue to add to the cost of running a large factory.

Some examples of how these two information and knowledge capabilities can be used to help improve operational efficiency are illustrated below.

Operational Modeling

OM&S is used widely in process development, wafer fabrication, assembly test, manufacturing support, and other parts of the manufacturing enterprise. Savings accrued through the use of OM&S can be substantial, in the hundreds of millions of dollars.

Generally speaking, OM&S capabilities are directly linked to improvement of major factory performance metrics: cost reduction, delivery improvement, quality improvement or product performance improvement. Factory improvement issues are often stated thus: "If I change this and that, how does the result affect my bottom line performance?" or "What if I did this instead of that (if I added or removed people from the line; if I laid out the equipment differently; if I used this strategy vs. that one to schedule downtime, and so on), how would factory performance be changed?" Consequently, OM&S programs are often called "what-if" scenarios. They are used to save time and money. Running a physical experiment, i.e., re-laying out a product line, can take months or years compared to running a simulation, which can take minutes or hours. Or, running a physical experiment can cost too much. Running a single experiment in an operating fab could cost hundreds of thousands of dollars.

Application of OM&S Technology

The following are examples of how OM&S technology can be applied:

- Comparison of Continuous Flow Manufacturing (CFM) to current Functional Flow Manufacturing practices in the production of Single Edge Connector Cartridge (SECc) modules may be applicable to other manufacturing facilities.
- Dedication of particular stepper lenses to particular lots in fabs to improve overall factory performance.
- Increase in WIP turns using full factory simulation to enhance use of information to improve performance.
- Evaluation of the effects of lot size on factory performance to determine optimum lot size.
- Evaluation of the effects of modifying operational policies on scheduling use of factory equipment to increase utilization without adding more equipment.

More detailed discussions of applications of operational modeling may be found in Court Hilton's paper entitled "Manufacturing Operations System Design and Analysis" and Karl Kempf's paper "Improving Throughput Across the Factory Life-Cycle" also appearing in this Q4'98 issue of the *Intel Technology Journal.*

Note that all of the above examples are specific applications; they do something for someone who has a specific issue to resolve. As such, they are highly beneficial. But the real payoff comes when all these applications are linked through some integrated, hierarchical model. The benefits of such a model can be imagined by comparing it to Microsoft Windows*. In Microsoft Windows, each application (Word*, Excel*, PowerPoint*, etc.) is individually very useful, but the ability to share textual and image objects between applications greatly enhances the whole. The total Windows environment is more than just the sum of its parts.

So part of the evolving OM&S effort is aimed at defining a modeling hierarchy, and establishing the links and infrastructure between modeling elements, to make the entire modeling environment much more than the sum of the individual components. This is schematically illustrated in Figure 4, where the NOW environment shows individual models, distributed through the manufacturing enterprise, and the FUTURE scenario shows an evenly distributed, linked hierarchy of models.

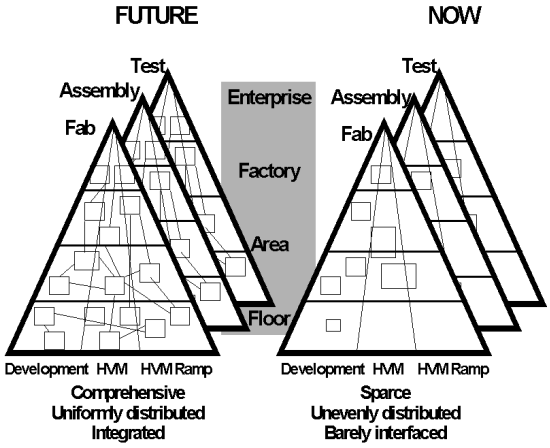

Figure 4: Modeling hierarchy

The scope of operational modeling is very broad, as illustrated in Figure 5. For convenience, the operational environment has been divided into three roughly equal domains: those dealing *directly* with product (the PHYSICAL DOMAIN), those dealing with the data and information associated both with the product and with the factory itself (the INFORMATION DOMAIN), and those dealing with background and support issues (the INFRASTRUCTURE DOMAIN). Each of these domains is itself sub-categorized, as shown in Figure 5.

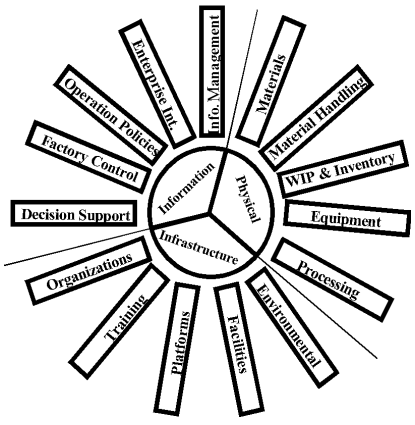

Figure 5: Model scope

Each sub-category is made up of sub-sub-categories, and so on, until one reaches the lowest level of the model hierarchy. Hence, each topic can have applications, roadmaps, goals, interfaces, etc.; the question is, how many of these topics

* Other brands and names are the property of their respective owners.

have common elements and should actually be integrated with one another. This integration is both lateral, meaning across equivalent levels of hierarchy, as well as being up and down the chain of model hierarchy. It raises interesting philosophical questions about model integration, as well as deep practical questions of how one may make modeling capabilities more cost-effective and efficient.

Knowledge Management

Whereas OM&S technology provides a fairly direct link between the capability of a technology and factory performance, knowledge management (KM) technology is one step removed from such a direct link. Indeed, KM is a logical counterpart of physical asset management, the leveraging of our physical capital (land, factories, computers, equipment, etc.) to improve profitability. KM leverages "knowledge capital" (patents, trademarks, know-how, competencies, skills, tacit or unwritten knowledge, relationships, etc.). Since, at the present time, the value of these intellectual assets is not really understood, the first goal of KM is to define a set of metrics that allows one to know even if there is any leverage to intellectual capital.

One rough estimate may be made by comparing the value of a company in the eyes of its stockholders to the paper value of the company's physical assets. In the case of Intel, the stock value (shares outstanding times price) is about $120 Billion, while the physical assets have a value of about $25 Billion. The difference, about $95 Billion, or four times the physical asset value, may be ascribed to non-physical assets!

KM capabilities may be defined using the following model. KM is divided into four large domains: the creation of knowledge, the capture and structure of knowledge, the dissemination of knowledge, and the application of knowledge. Some attributes of each of these four categories are shown below in Table 5.

The two areas that require most attention are items 2 and 3 in Table 5: the collection, structuring, and indexing of knowledge, and the secure, rapid dissemination of knowledge to potential users. Of primary interest are metrics: understanding how to value the intellectual assets of the enterprise, and indexing: the categorization of knowledge for rapid and ubiquitous application. Also of great significance is the knowledge tool environment. Much like the information tools of prior generations, knowledge tools are rapidly emerging and evolving. We expect that a knowledge tool environment similar in concept to the Windows* information environment will emerge, thereby allowing us to exchange knowledge objects in much the same way as we already exchange information objects.

1. Knowledge Creation
- Research
- Brainstorming
- Strategizing
- Synthesizing
2. Knowledge Structure
- Data and knowledge databases
- Indexing
- Training development
- Report generation
- Knowledge management tools
3. Knowledge Dissemination
- Inter- and Intranet
- Education and training
- Electronic mail
- Reading
- Browsers and interfaces
- Security precautions
4. Knowledge Application
- Problem solving
- Strategizing
- Decision making
- Managing and metrics

Table 5: Knowledge management domain

Some potential areas where knowledge management can be applied are as follows:

- Understanding and matching of core competencies of individuals with attributes of job needs.
- Providing a "Knowledge Atlas," a visual environment in which employees can guide themselves to find knowledge items, for example, "how do I do this?"; "who do I see to do that?"; or "who is the expert on this?".
- Developing tools that leverage an employee's job skills, allowing people to take on more responsible jobs using knowledge assistants for help.
- Better problem solving by providing access to vast and comprehensive knowledge bases of past occurrences, tied to the nature of a problem rather than to simple keyword searches.

Knowledge management tools will help make us a more efficient company by providing access to knowledge to people who need it, wherever they are and whatever the problem set. We should then be able to make faster and wiser decisions, resulting in significant improvements in factory and even enterprise performance.

Organizational Issues

Pursuit of information and knowledge technology, as given in the examples above, is not free. In particular, in addition to the obvious need for technical skills, there is a need to understand and respond to the managerial and organizational skills required for success.

At one time, the resources required to operate a factory consisted almost universally of people who had their hands on the product: moving it, processing it, assembling it, storing it. Currently, the trend is towards having a greater percentage of the workforce spending time on the processing of data and information. They gather data, analyze data, and convert these data to information. This information is then stored, transmitted, and disseminated, so that decisions can be made and our knowledge increased. Meanwhile, the total workforce is decreasing through physical and logical productivity improvement.

The result of these two trends is schematically illustrated in Figure 5 below. The total workforce is decreasing, while the percentage of IT and software personnel is increasing.

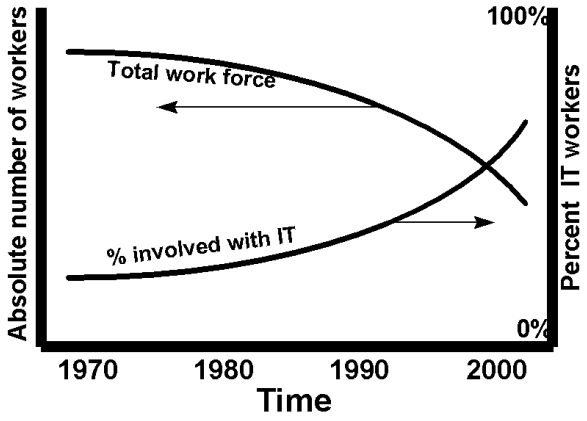

Figure 6: IT headcount projections

There are two personnel issues to confront as a result of these trends: the first is the evolution of the factory workforce from process-centric to one that is more information-centric. The processing domain is equipment dominated, where our equipment suppliers own the core competencies. As more and more information processing is incorporated into the factory, more technologists will be necessary in the IT processing field. However, this problem is fairly manageable; Intel is an expert at managing technology.

The real issues are those of organization and management. Managing process is straightforward: align the management organizations functionally, for example, with cross-cutting metrics such as yield, cost, delivery, etc. Managing the information organization is different, however. The cross-cutting disciplines such as platforms, software, and databases are not conducive to factory management, but the information technology does not map well to the traditional metrics of yield, delivery, etc. Furthermore, the skills of management need to be different. Management needs to be more proficient in IT skills; their current skill set is technologically oriented towards processing technology.

These management and organizational issues need to be dealt with concurrent with the growth of IT technology.

Conclusions

It seems clear that our industry is departing from at least some of Gordon Moore's earlier quantitative predictions. One of these is illustrated in Figure 6. Gordon's 1974 tongue-in-cheek but genuine extrapolation of wafer sizes suggested that by the year 2000, we would have 57-inch diameter wafers! Clearly, this is off by about an order of magnitude. Yet simple extrapolations of Gordon's trends does lead to qualitatively correct predictions.

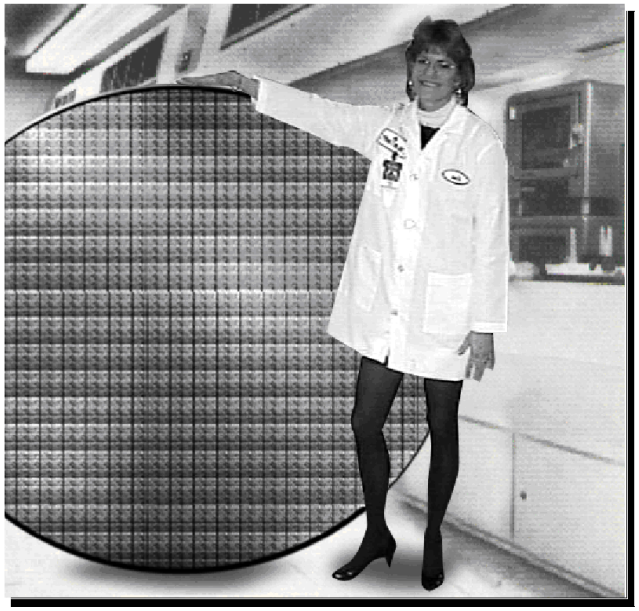

Figure 7: "Extrapolated" Year 1999 wafer size[1]

Regardless, two trends seem inescapable: everything in the production of semiconductor devices is moving toward more expensive factories, and there is swiftly expanding use of information and knowledge to reduce costs, improve delivery, and improve quality. These two trends need to be linked to try to alleviate the effects of the former by using the latter. At the same time, one must also recognize the emergence of other forces: the need for cleaner, safer, and less energy-consuming manufacturing enterprises, the evolution and indeed revolution of materials and materials' processing, and

the change from local politics and culture to global politics and culture. All these trends will result in a significantly greater emphasis being put on manufacturing as a competitive weapon in the 21st century.

Acknowledgments

I would like to thank Gordon Moore, Karl Kempf, Court Hilton, Sri Sridharan, Scot Ruska, Bruce Sohn and John Birchak for valuable discussions that helped formulate the concepts discussed in this paper.

References

1. Gordon Moore, *Electronic Materials Symposium* Santa Clara, CA, March 1998.
2. "Next Generation Manufacturing: A Framework for Action," *Agility Forum*, Bethlehem, PA, 1997.

Author's Biography

Gene Meieran received his B.S. degree from Purdue University in 1959 and his Sc.D degree in materials science from MIT in 1963. He joined Fairchild R&D in 1963, where he specialized in the analysis and characterization of semiconductor device materials. He joined Intel in 1973 as manager of Package Development, responsible for developing new lines of packages for the emerging memory and microprocessor products. In 1977, he joined the Quality and Reliability staff, with responsibility for all Intel materials, the Materials Analysis Laboratory, and for manufacturing reliability functions. He has worked in Statistical Process Control (SPC) and advanced manufacturing strategy development in Intel's Technology Manufacturing Engineering group for the past 12 years.

Dr. Meieran taught technical courses in leading US universities and has given seminars and invited talks to many international universities. He has about 50 technical awards and has received three international awards based on technical talks.

He served on the Scientific/Education Advisory Board for Lawrence Berkeley Labs and on advisory boards for several university departments. He has been Director for Research for the MIT Leaders For Manufacturing Program since 1993 and has served on numerous government and industry panels dealing with manufacturing technology and policy issues.

In 1985, Gene was appointed an Intel Fellow, Intel's most senior technical position. In 1987, Purdue University elected him a Distinguished Engineering Alumni, and in 1998, he was elected to the National Academy of Engineering.

His e-mail is gene.s.meieran@intel.com.

The Evolution of Intel's Copy EXACTLY! Technology Transfer Method

Chris J. McDonald, Intel SEMATECH

Index words: copy exactly, technology transfer

Abstract

Semiconductor manufacturing is characterized by very complex process flows made up of individual process steps, many of which are built to very close tolerances. Furthermore, there are complex interactions in these process flows, whereby each process step can affect many other steps, and each final device parameter might be determined by the results from many inputs. This level of complexity is increasing with each new technology generation. Items that were once considered second-order effects, such as barometric pressure and ultra pure water temperature, are now important variables affecting process results.

The costs of technology development and capital equipment for production are very high and are increasing with each generation, thus making technology transfer very important. Once a new process flow and product portfolio have been developed, it is essential that the technology transfer to mass production take place as quickly as possible, without disruptive quality issues, and with the highest possible yield. No time is available to debug new problems that occur during the transfer.

The traditional technology transfer approach often allows many equipment and process changes to be made. These changes are intended as improvements in the process, or they are for the convenience of the production factory, which may be already producing other products. As semiconductor technology becomes more complex, these changes have resulted in unforeseen problems that cause production start-up delays and inferior results.

The Copy EXACTLY! philosophy and systems were developed [6] in order to minimize the time required for a technology to be transferred and to ensure product quality and yields are not compromised. The methodology has been improved and refined, and has become an important element in Intel's overall manufacturing strategy [1]. This paper describes the Copy EXACTLY! methodology and the increase in technology transfer performance that it has brought about. Some side benefits of this methodology are also discussed.

Introduction

Table 1 shows the typical technology transfer approaches used over the last ten years or so. At the 1.5-micron generation, process flows were much simpler than they are today. A small band of technical experts would typically be employed to orchestrate a successful technology transfer. Generally there would be few ground rules. Since there is always a lengthy "certification" or "qualification" exercise to prove product quality and reliability, the transfer from R&D to manufacturing, or to a new factory, offered the opportunity to introduce improvements to the equipment and process. The latest model equipment or even a new vendor might be chosen. Process recipes could be changed to improve them. In the case of an existing factory picking up a new process flow, changes were made to match existing processes and methods to improve efficiency and productivity. Sometimes, a wafer size conversion would even be made at the same time, involving many changes. Overall, however, the number of variables was relatively small, which made it simple to trouble shoot any results that did not come out as expected.

Technology Generation	Transfer Strategy	Comments
1.5 micron	"Make It Work"	Small band of engineers. Few ground rules needed.
1.0 and 0.8 micron	"Process Output Matching"	Copy selectively. Match to existing factory conditions.
0.5 micron	"Copy EXACTLY!"	Copy everything that might affect the process.
0.35 and 0.25 micron	"Systems Synergy"	Copy all manufacturing systems.

Table 1: Technology transfer strategies

For the one-micron generation, technology transfer started to get more complicated. A structured methodology was needed, whereby each process step would be measured to ensure it matched a target value or complied with a set of specification limits [2] [3]. Most projects, however, only fo-

cused on matching device and final product parameters. As long as these were correct, changes would still be introduced as a part of the transfer process.

For the sub-micron generation, the above approach has had its share of problems. There are many more process steps today, and many of them are made up of several components. For example, a typical metal or dielectric layer is now a sandwich of multiple layers of different materials and compositions. Very fine device structures are subject to different effects, such as inter-layer stresses and adhesion. Phenomena that were once considered second-order effects now have a significant effect on the process result. Among these are barometric pressure and ultra pure rinse water temperature. In general, the process is manufactured with much closer tolerances, increasing the importance of process control. Even the length of an electrode cooling hose has had a catastrophic effect, but this is a very subtle variable to find. With larger die sizes, defect control becomes even more critical, and the way the process is actually run becomes a more important factor. An example is preventive maintenance intervals and workmanship. Many of the factors just mentioned are difficult to measure and quantify, which makes them dangerous "unknowns" during a technology transfer project. When many changes are made, the risk of something going wrong is greatly increased. Moreover, if something does not come out as expected, the number of variables that have to be studied when trouble shooting the problem is greatly increased. The amount of experimentation and therefore, time, required to find the problem increases as a power function of the number of variables involved. If the problem is a showstopper, for example affecting the product reliability, the end result is a costly delay. Even if this is not the case, yields may be depressed for an extended period.

Figure 1 shows one example of using the traditional approach for the 1.0-micron technology generation. The die per wafer yield is one of the most important variables in wafer fabrication and is used in this example. (The graph is normalized for die size). Other factors affecting product performance or manufacturing efficiency showed similar trends. The first production factory in this example, which was a brand new facility, obtained results that reasonably matched the parent R&D line. You can see the yields improve further as improvements were made and the organizations moved down their learning curves. Note, however, the divergence at the end of Year 2. The technology transfer results were good [2]; however, yields diverged as the R&D line focussed on yield improvement, and the manufacturing line concentrated on increasing volume.

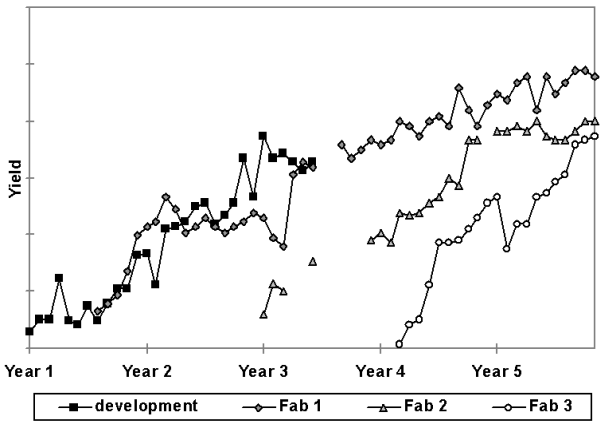

Figure 1: Traditional technology transfer method for the 1.0-micron generation

Eventually the yields did converge. The second and third factories, which were already manufacturing different process flows, made many changes to the process and equipment and used their existing manufacturing methods. It took several months of task force actions to catch up, by which time the first factory was moving further ahead. In effect, the same learning curve was repeated independently by every factory. For the 0.8-micron generation, a new factory start up and a new product introduction were delayed by three months while a device reliability problem was solved, and it took approximately one year to obtain equivalent yields [4].

Method

For sub-micron technology, it was realized that a fundamentally new approach would be needed in order to accomplish an "order of magnitude improvement" in the effectiveness of technology transfer. The Copy EXACTLY! philosophy and systems were developed [5] [6] for the 0.5-micron generation, and they have become a key part of Intel's manufacturing strategy [1]. The capital letters, underline, and exclamation point emphasize the paradigm shift that is required to transfer technology using this method.

Copy EXACTLY! Philosophy

Stated in its simplest form, "everything which might affect the process, or how it is run" is to be copied down to the finest detail, unless it is either *physically_impossible* to do so, or there is an *overwhelming_competitive benefit* to introducing a change. This philosophy differs greatly from the traditional method. In practice, there are always some issues

that crop up and prevent an exact copy being made, so it was important to provide for making some changes in a controlled fashion without opening the field too much. For example, in Europe the supply voltage and frequency can be different than those in the U.S., so these had to be accommodated. Moreover, engineers are typically trained and rewarded for making improvements, which in the semiconductor industry implies orchestrating change. Even the educational system stresses independent work, and copying is seen as cheating. Making a philosophical statement is obviously much easier than implementing it within a large team of R&D and manufacturing engineers. Therefore, a comprehensive set of systems was put in place to ensure it would be implemented, and this set of systems is discussed in the next section.

Systems

The systems that were implemented are as follows:

Four level matching: Traditionally, it has been considered acceptable if the final product parameters are matched between the R&D line and manufacturing. However, the Copy EXACTLY! approach requires four levels of matching. These are illustrated in Figure 2.

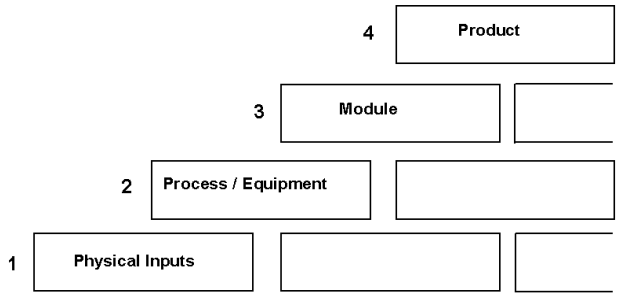

Figure 2: Four-level matching

- Firstly, the physical inputs have to be matched. These are the energies and materials supplied to the process chambers: for example, gas flows, temperatures, pressures, RF power, and so forth. These might be supplied to the equipment by external sources or be generated within the equipment itself. Everything about the equipment and its installation must be an exact copy down to the diameters of piping and the number of bends, board revisions, software, etc. The settings for these parameters and anything that might affect them are copied. Standards are generated to allow measurement and comparison, and the values are measured and matched.
- Secondly, data is collected at the process step output level on parameters such as film thickness, resistance, composition, etc., and they are compared to results at the R&D site.
- Thirdly, a comparison is made at the module level, using test structures such as oxide capacitors and metal serpentine patterns.
- Fourthly, the actual product characteristics are measured and matched.

Formal statistical tests are used at each level. If the match passes these tests, then we proceed to the next level and so on. If the match does not pass the tests, the root cause must be found and eliminated. If it can't be found, trouble shooting occurs to find out which of the previous level inputs is responsible because, despite best efforts, something may have been overlooked. It is vitally important to avoid the temptation to make a compensating adjustment. Due to the complexities involved, an adjustment may result in an interacting parameter, possibly something not measured, being mismatched.

A change control system: Most factories have some kind of approval process for making changes to a production process, either in the form of a sign-off list or a formal change-control committee. Generally there is some kind of record of the data showing the benefits of the change. The R&D line continues to make improvements to finish off the technology development and, in many cases, they may also run some level of samples and production output. With Copy EXACTLY! change control is started before technology transfer, and all changes are implemented directly into both the R&D and production lines within one week, or according to an approved schedule. The pace of R&D work is not allowed to slow, so careful planning is required to ensure the new line is ready to accept the changes in real time. Any engineer from the manufacturing line who has a good idea for improvement is encouraged to pursue it. The only difference from the traditional approach is that the idea must be implemented simultaneously at all sites. The change control board is responsible for the smooth operation of the system, which includes ensuring that the additional requirements do not slow down the rate of improvement.

Equipment difference form: In the Copy EXACTLY! system, each first piece of equipment in the new factory or on the new process flow in the existing factory is treated as a change, subject to change control. Audits are conducted and an Equipment Difference Form is prepared from each. This form documents the actual difference, what risks it might pose, and the corrective action plan. This is formally reviewed by management.

Supplier education: Equipment and materials' suppliers are constantly improving their products in response to demands from the semiconductor industry for improvement. These changes are still desirable; however, with Copy EXACTLY! they are first introduced into the R&D line and from there transferred to production. The suppliers are a vital part of the technology transfer and need to be thoroughly educated on the new concept and systems.

Audits: An audit is a formal procedure whereby engineers from R&D and from production audit both lines. These audits are required and scheduled as part of the technology transfer and are ongoing for a period thereafter. A report is written for each audit, detailing plans to correct all differences found.

Joint specifications: Since the equipment, process recipes, and procedures are all the same, there is no reason why the documents provided for training and manufacturing operations cannot be the same. These are not copies; they are the same documents, either paper or electronic.

Questions and answers (Q&A): Different engineers tend to interpret the Copy EXACTLY! message in different ways. For example, some engineers might say, "Surely if I make sure the pressure is the same, then it doesn't matter if I use a different pump with less bends in the vacuum line." The answer to this particular question is "*Yes,* it does make a difference, and *no,* it's not ok to make a change." The rationale is that you might be able to get the same result under ideal conditions, but the only way to guarantee you will always get the same results, both steady state and transients, under all possible conditions of environment, age, etc., is to copy the configuration exactly. To deal with this type of question, a detailed Q&A list was prepared and communicated to all engineers involved on the project.

Systems Synergy

The scope of copy EXACTLY! for the 0.5-micron technology was for the most part limited to anything that might have an impact on the process, or how it is run. The motivation was to guarantee equivalent yields, starting with the first wafer, and to ensure there were no reliability problems to delay production. One recommendation [7] from this program was that the concept could be extended into other areas as a way to further accelerate new factory start-ups and the manufacturing ramp of new-generation technologies.

The 0.35- and 0.25-micron generation technology transfers [8] took the Copy EXACTLY! method a step further into what has been described as total "Systems Synergy," where almost every aspect of the fabs are identical at multiple geographic sites. The 0.8- and 0.5-micron generations both had a "virtual factory" organization structure and a series of committees to set the strategic direction and manage the technology. For the 0.35- and 0.25-micron technology, this has been expanded: a Steering Committee at the plant manager level sets the overall direction, Joint Engineering Managers' Teams manage the technology, and individual Joint Engineering Teams work the details at the process and equipment level. Similar structures exist in other areas, such as Manufacturing Operations and Automation, and a "Joint Synergy Team" manages the overall system.

Results

Figure 3 shows some typical results obtained from Copy EXACTLY! Two new factories were successfully brought on line with the same yield results as the parent R&D line. Furthermore, all three lines were able to improve their yields together by implementing improvements simultaneously. Other parameters such as product quality and reliability and manufacturing efficiency also matched very closely.

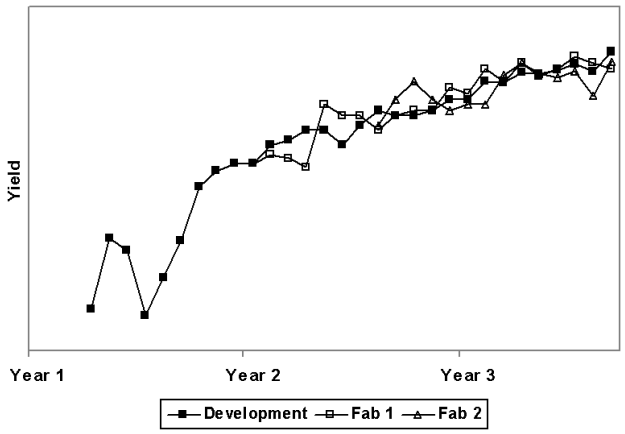

Figure 3: Copy EXACTLY! technology transfer method for the 0.5-micron generation

As always with projects of this magnitude and complexity, there were some issues encountered along the way. For example, a very subtle problem affecting the integrity of the sub-micron metal lines was found. However, since the process had been copied so precisely, trouble shooting became an exercise in revisiting the exceptions that had been made and auditing to look for unforeseen errors in copying. In this example, two variables were identified as suspects, and a simple experiment on test wafers identified the cause within a week. A simple typographical error had been made in entering a process recipe. The problem was very subtle and would have taken many weeks to identify if a traditional transfer approach had been used. In addition, areas for improvement

in the technology were known and found in both sites. Since no new problems were introduced as a result of the technology transfer, the number of engineers and other resources available for basic improvement work was greatly increased. Moreover, the overall technology transfers to two new 0.5-micron factories were accomplished in record time with very few problems along the way.

Figure 4 shows the results obtained on the 0.25-micron technology generation, using copy EXACTLY! and Systems Synergy.

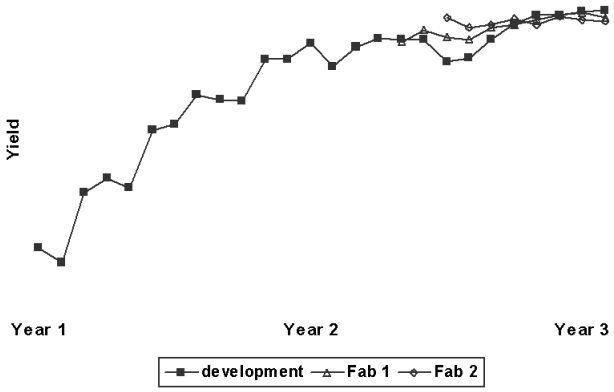

Figure 4: Copy EXACTLY! technology transfer method for the 0.25-micron generation

It is now routine for new Intel factories or new technologies that were transferred to obtain equivalent yield starting with the first check-out wafer. Production quantities of products are started immediately since there is such confidence the product will be good.

Discussion

The difficulties in implementing this new philosophy and system are not to be underestimated. Any major project, such as a new factory or new process flow in an existing factory, is started with the ambition to get the best ever results. Engineers are trained and rewarded for doing improvement projects, and product re-qualification affords them such an opportunity. The natural tendency is thus to use the new start-up as an opportunity to implement improvements. To change the mind set of a large body of technical experts requires a very simple message, consistently delivered, and backed up by a set of systems that make it difficult to behave differently from the desired state. With Copy EXACTLY!, the message to the production engineers was to achieve the best ever replication in the fastest possible time, and it will be considered the "best ever." Once the new products are up and running, with good stability and in high volume, the production line engineers earn the right to take a leadership role in making improvements. In the meantime, ideas are still welcome, but they are implemented through the R&D organization and at the same time at both sites. The R&D engineers also need to make some sacrifices. To make changes they now need the support of the production line engineers.

The results obtained clearly show the merit of the Copy EXACTLY! philosophy and systems. The process flow was transferred to two new factories in record time with equivalent yield and other indicators, and with no product quality issues. The new lines were able to precisely intercept the technology learning curve.

A number of other benefits were also realized and are as follows:

Customer acceptance: Many major customers for integrated circuits are well aware of the risks in changing manufacturing plants and will typically demand the opportunity to re-qualify a second source. If the supplier has a high credibility rating, this may simply require a study of all the data from the new line. However, in many cases the customer may want sample devices to submit to his/her own testing, or to a third party laboratory. He/she may also require a site visit and an opportunity to audit the new line. In all cases there will be additional costs and delays in time to market. However, once the customers understood the Copy EXACTLY! method, many of these concerns, costs, and delays were eliminated.

On-going mutual synergy and shared learning: In the example outlined here, the R&D line continued to manufacture the new products along with the two manufacturing lines. By keeping the process in lock step at all three sites it was possible to share the improvement projects among them. Improvements were characterized in one site and transferred to the others with minimal effort. In effect, the number of engineers per process step or per area for improvement is increased, as is the number of improvement ideas generated.

Manufacturing flexibility: With three sites running the exact same process, products were easily transferred back and forth with no re-qualification, other than checking the mask set. Using free capacity at another site has also solved manufacturing bottlenecks.

Conclusion

The Copy EXACTLY! method has proven itself as a technique for semiconductor technology transfer. A new process flow and products can be introduced to production in minimum time with equivalent yields and without the introduction of product-quality issues. Both manufacturer and customers can reduce their time to market. This approach could equally be employed in other industries where the technology is complex and has many interacting variables affecting

the end result. The concept has been successfully expanded to cover all systems used in manufacturing.

Caution

Copy EXACTLY! is a powerful method for technology transfer, but should not be applied during technology development. By definition, technology development means taking new processes and improving and integrating them to create a new generation process flow with greater capabilities. While it may be decided that some existing process modules and equipment can be reused as they are, in general, technology development requires great creativity and innovation. Technology development would be dampened by the rigid discipline required during technology transfer and manufacturing.

High-volume manufacturing also demands a high degree of change. Yields must be continuously increased, efficiencies improved, and costs reduced. It is vitally important that the systems used to manage change strike a good balance between the discipline required to keep the factory under control and the creativity and innovation required for continuous improvement. Manufacturing improvement systems need to be very fast-moving and flexible. Multiple factories running the same process and products should remain matched, but not necessarily identical at all times. New approaches should be tested in one site and proliferated to others when proven.

Finally, the Copy EXACTLY! method is designed to match all factors that impact the process or how it is run. Other systems might benefit from matching, but time and money should not be wasted on matching factors that have no impact on the overall process.

Acknowledgments

The author would like to recognize Tom Hartman for his leadership through two new factory start-ups. The author also acknowledges Bob Jecmen and Bruce Leising for their sponsorship in the implementation of these new approaches, and for their support in the publication of this paper.

References

[1] G.H. Parker, "Intel's Manufacturing Strategy," presented at the *5th International Symposium on Semiconductor Manufacturing.*

[2] C.J. McDonald, "Fab 9.1 Start-Up Methodology." *Intel Internal Document,* June 1988.

[3] M. Bohr and S.Ruska, "P650 Transfer Methodology Proposal," *Intel Internal Document*, December 1989.

[4] S. Ruska, "Post-Mortem of Fab 9.2 Yield/Undercut Issues," *Intel Internal Document*, December 1991.

[5] C.J. McDonald, "Fab 10 Start-Up Methodology – Copy EXACTLY!," *Intel Internal Document,* February 1992.

[6] C.J. McDonald, "Copy EXACTLY! A Paradigm Shift in Technology Transfer Method." Presented at the *IEEE Advanced Semiconductor Manufacturing Conference*, 1997.

[7] J. Multani, C.J. McDonald et al., "P852 Virtual Factory Vision Through Copy EXACTLY!." Presented at the *Intel iMEC Conference*, 1994.

[8] G. Gimpelson et al., "Rapid Technology Transfer: BKM's from the P856 Process Transfer." Presented at the *Intel iMEC Conference*, 1998.

Author's Biography

Chris J. McDonald holds a B.Sc. in electronics from the University of Southampton, UK and an M.Sc. in electrical engineering from the National Technological University, USA. He is a sixteen-year employee of Intel, currently assigned to SEMATECH where he is director of the Advanced Technology Development Facility. His other positions included that of manufacturing manager and yield/Q&R manager for the start-up and ramp of Intel's high-volume manufacturing facility in Ireland and a process/equipment engineering manager in Albuquerque. He has twenty-four years experience in semiconductor process development and manufacturing, and he has worked in all areas of wafer fab engineering with a wide range of memory, microprocessor, and other products. His e-mail is chris.mcdonald@sematech.org or chris.mcdonald@intel.com.

Manufacturing Operations System Design and Analysis

C. Hilton
Manufacturing Strategic Support
Technology and Manufacturing Engineering, Intel Corp.

Index words: factory, manufacturing, discrete-event, simulation, modeling

Abstract

This paper describes manufacturing operations design and analysis at Intel. The complexities and forces of both the market and the manufacturing process combine to make the development of improved semiconductor fabrication manufacturing strategies (like lot dispatching, micro and macro scheduling policies, labor utilization, layout, etc.) particularly important. We present modeling as an effective way to further this improvement. We describe various categories of models and why they are useful. We present an overview of eight examples of how we are using modeling to improve manufacturing performance and cost. These summaries illustrate how millions of dollars have been saved in direct cost and/or cost avoidance. We conclude with a brief discussion of next steps and cautions for those establishing a manufacturing operations' group.

Introduction

In *A Tale of Two Cities*, Charles Dickens begins, "It was the best of times, it was the worst of times…" [1]. Never in history has more reliable semiconductor power been available to the consumer at such a low price. And never in history have the pressures on the manufacturers of these devices been more severe. In essence, it is the best and worst of times. Product complexity is rising, market and product segmentation is increasingly fracturing the market, lead times are shrinking, available margins are shrinking, the cost to bring a semiconductor fab on line is doubling every three to six years, and the historical avenues of cost improvement—device yields, line yields, and even device shrinking—are all approaching limits.

Intel's product manufacturing occurs in several phases: fabrication of the silicon-based device, testing, assembly and packing of the device, testing again, and sometimes assembly of the computer system or sub-system together with its testing yet again. Further complexity is introduced into each of these manufacturing processes in an effort to meet the demands for special market-driven product features.

The wafer fabs, which produce the silicon-based devices, use complex processes involving two billion dollars of equipment and 300-500 operational steps all housed in an ultra-clean environment. Fabs typically require on the order of a thousand direct staff to operate. The material does not move through the factory in a linear fashion from front to back, but loops back on itself, revisiting some areas sometimes more than twenty times. This creates all sorts of "feedback loops" (known as re-entrancy) in the dynamic response of the factory to perturbations on the factory floor and in the marketplace. These perturbations can be dampened or amplified depending on factory design, operational policies, and the current state of the material being manufactured. The majority of wafer fab cost arises from capital equipment costs. Wafer fabs feed other assembly and test operations where the devices are packaged and tested under a variety of additional constraints. Assembly and test manufacturing flows are generally linear and have several dozen processing steps. To build and equip a factory costs in the order of hundreds of millions of dollars. These factories require hundreds of direct staff to operate.

Most of the high-revenue products being manufactured today did not exist two years ago. Market forces, often not well understood, can drive product functionality as well as product packaging through radical changes within a short period of time. Each new generation of product, introduced every 12-24 months, requires new and even more costly process equipment. Often this equipment is itself on the cutting edge of technology and does not always have the performance or reliability desired for cost-effective operation. This short product lifecycle makes the reuse of equipment and the flexibility of factory use very important.

At the SEMI-Advanced Semiconductor Manufacturing Conference in 1997, Clark Fuhs, Director/Principle Analyst of the Semiconductor Manufacturing Group for Dataquest said, "The next productivity leap in the semiconductor industry will have to come through the implementation of manufacturing science and of industrial engineering practices" [2]. Manufacturing science, or "factory physics" as we like to refer to it, "is a systematic description of the underlying behavior of manufacturing systems" [3]. To understand this underlying behavior of systems of this complexity and cost, we typically do not experiment directly with the manufacturing operation. Rather, validated computer-based models are built to describe the behavior, and these models are used to develop and test factory design and operational practices that will optimize factory performance and flexibility at the lowest cost.

Models: Lies Or Oracles?

From a certain point of view, a model is a lie. From another perspective, it may be an oracle. A model is a lie in the sense that it is a purposeful simplification of a problem with the intent of focusing attention on what are believed to be the critical-few discriminating attributes or salient concepts. A model can be an oracle in the sense that an appropriate model allows us to conveniently manipulate complex systems and find answers to questions we could not approach in any other way. Models may be very detailed or very general (i.e., they are written at different levels of abstraction). Various model evaluation techniques are used within computer-based models, from simple spreadsheets to complex full-factory discrete-event simulations. Table 1 is a comparison of various types of model abstractions used in a study of a representative assembly plant operation. [4]. Note in Table 1, MTTF and MTTR refer to equipment Mean- Time-To-Fail and Mean-Time-To-Repair.

This paper focuses primarily on describing applications of computer-based models that attempt to represent complex and dynamic manufacturing interrelationships. This class of models is known as discrete-event simulations (DES). DES models are used at Intel to evaluate a variety of system performance, design, automation, and operational issues in a cost-effective, non-disruptive, statistical, and realistic fashion. Figure 1 shows the 12 areas in which we focus our efforts.

DES models can often help in decision making whenever one or more of the following conditions exists [5]:

- equipment utilization is greater than ~80%
- synchronization or merging of separate operations occurs
- manufacturing actions interact with outside events
- operations with widely variable completion times interact
- contention for resources or timing constraints exists

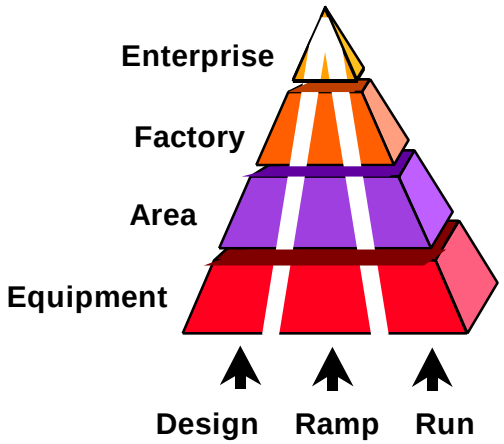

Figure 1: Scope of operations modeling activities

Analytic Method (model abstraction)	Level	Assumptions	Accuracy
Static algebraic analysis of linked systems without buffering	System	No buffering allowed Simple MTTF & MTTR model	Good model, Averages
Linked static analysis with buffering	System	Buffering allowed Simple MTTF & MTTR model	Better model, Averages
Monte Carlo analysis of static models	System Area	Distributions of input parameters are available	Better model, variability
Dynamic discrete-event simulation analysis	System Area Equip.	Buffering Failure & repair distributions	Best model, dynamic, variability
State-space analysis	Equipment	Accurate observation Fast transitions	Best validity (i.e., nature)

Table 1: Example comparison of various types of model abstractions used in solving an equipment linking design problem for an assembly operation

The Design of Experiments

Infinite Possibilities

A number of designed experiments, using static and discrete-event simulation models, are required to determine the expected factory performance under various conditions. To limit the infinite number of possible experiments, fractional and full-factorial designs can be iteratively used to define efficient sets of factory conditions to simulate.

Inputs and outputs

Our simulated factory inputs and outputs are similar to those found in an actual factory. Inputs include equipment count and layout; process time (units per hour); distributions for mean time to fail, repair, and assist; preventive maintenance (PM) time profiles; work in process (WIP) management policies including materials release; setup and batching policies; transportation times; labor availability profiles; and process flow definitions. Likewise, model outputs include equipment utilization, throughput times (TPT), and factory output (often referred to as "outs").

How Outputs are Reported

Simulation outputs are normally reported as differences or ratios of one scenario versus a baseline scenario. For example, rather than report a TPT, one would instead report that the simulation has a TPT that is 2.3 times the theoretical TPT, or that scenario B provides 10 percent more capacity than a baseline scenario A. This type of comparative reporting is useful because it allows the experimenter to focus on the key differences under study while normalizing away model simplifications in areas of less interest. For example, in this report we employ a "TPT Ratio," which is the ratio of a scenario of the simulated TPT, to an arbitrary baseline-theoretic TPT. The intent is for the ratio to be interpreted conceptually as a TPT.

Insight Through Sensitivity Studies

In a sensitivity study, outputs are measured over a range of input variable values. This is an excellent technique for understanding the expected range of the modeled factory performance, for developing intuition about how a specific manufacturing floor will perform, for understanding the impact on the study of questionable input data, and for verifying model performance. We have made extensive use of sensitivity studies in the majority of work reported here.

Example Model-Based Manufacturing Studies

Ergonomic Simulations

As unit volumes increase in the back-end, and as wafer size and weight increase in the front-end, our factory personnel experience ever increasing physical demands. In an industry unaccustomed to dealing with heavy and/or sustained physical labor, this poses new challenges for factory operation and design. We applied a specialized simulation-modeling environment that combines advanced software and computational techniques with standard ergonomic metrics and detailed full-motion models of humans [8]. The DES models, which define task type and frequency, are seamlessly merged with ergo models, which evaluate the physical impact of programmed actions on the worker. To enable visualization of the stresses, a display environment capable of stereoscopic 3-D showed worker actions in real-time from any perspective (including that of the wafer if you are so inclined!). In this environment, the upper limbs of virtual workers dynamically changed color to express the degree of strain to which they were subject during their activities. Ergonomic evaluations included reachability, field of view, RULA (rapid upper limb assessment), posture analysis, NIOSH lifting guidelines, energy expenditure, and activity timing. These models, together with other factory, cost, and strategic models were used to support our selection of an optimized wafer-lot size of 25 wafers to be used in our next generation 300mm wafer fabs.

Design for Environment

Intel is active in assuring safe and high-quality environmental conditions for its workers and the families that live in the communities where it operates. In addition, the environmental permitting process, required by our operations, often requires more lead time than it does to build and start up a manufacturing facility.

Design for Environment is integrated into our process introduction business plans and has targets that must be met, just like other process goals. Key to a Design for Environment program is the use of a variety of models including an integrated mass and energy balance model. This static model incorporates the best data and models from throughout the corporation and is interfaced with our factory design models. These models allow us to project, years in advance, the effluent and energy demands of our processes. Where needed, process changes can be made to assure environmental quality. This also allows us to effectively target R&D efforts with suppliers, universities, and national labs to assure more benign processes.

This approach is working successfully. As shown in Figures 2 and 3 we continue to use less water / (silicon area manufactured), and our air emissions are better from generation to generation.

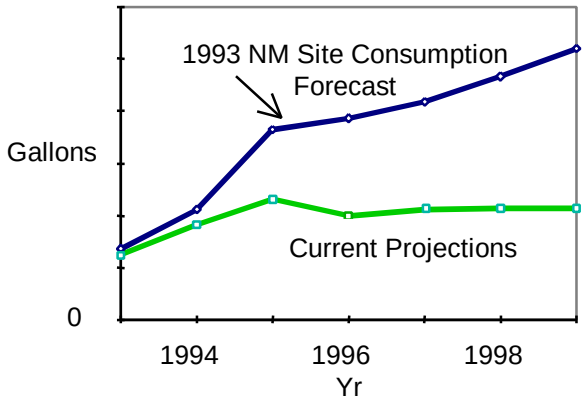

Figure 2: NM site water consumption per amount of silicon manufactured

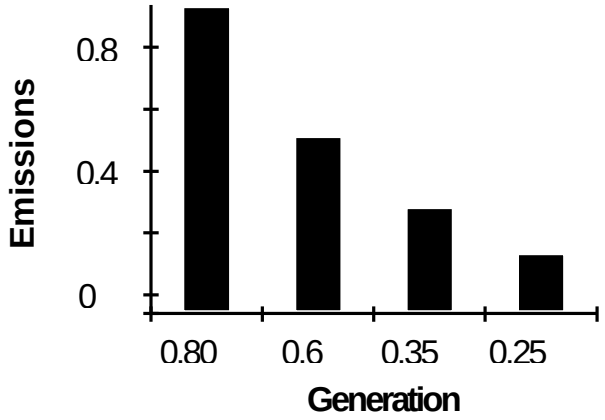

Figure 3: Targeted VOC emissions as a function of process generation (microns). Emissions are normalized by square centimeter of silicon devices manufactured.

We also have comprehensive programs to reduce, recover, and/or recycle a variety of our processing chemicals. Modeling also plays a role in these efforts. For example, one of the difficult challenges we face is dealing with the changes in the chemical content in the emissions from generation to generation. Our ability to effectively measure and model individual tools is key to determining the correct solution path. For example, in Figure 4 we can see that the total quantity and chemistry is changing for Per Flouro Compound (PFC) going from our 0.35 micron generation to our current predictions for the 0.13 micron generation by a factor of nearly 3X. In addition, the early generations contained relatively greater quantities of C_2F_6 and CF_4 when compared to the other effluent gases. The optimal technologies to deal with these kinds of emissions are fundamentally different. Models help us understand these requirements and assure appropriate technology is in place [6].

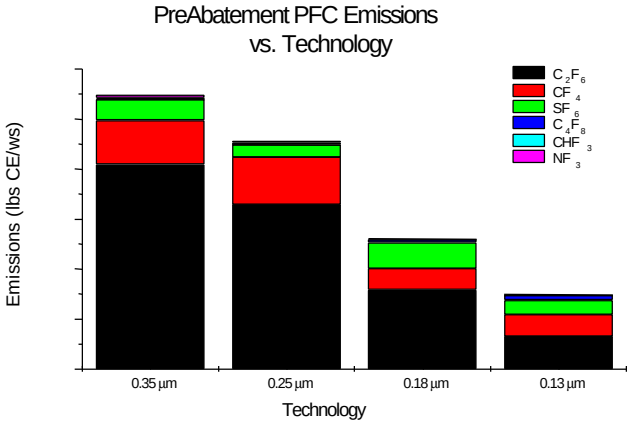

Figure 4: Pre-abatement PFC emmissions vs. technology

Assembly Floor Layout and Operation [7]

This project sought to integrate a variety of views about Pentium® processor assembly line layout from our manufacturing experts and evaluate performance of specific options as part of Intel's Copy EXACTLY! program to assure consistent implementation of our manufacturing operations policy across our plants worldwide. At one philosophical extreme is a flow layout that mimics the process flow, with process equipment hard-linked together for streamlined processing. This approach may be justified by the expectation of improved TPT and the immediate visibility provided to any equipment failures. At another extreme is the functional or grouped layout where like machines are grouped together on the floor, and product is routed to the first available machine. A justification for this approach is the expectation of improved capacity flexibility and robust performance in the presence of equipment service needs. In this study we evaluate these extremes and hybrids of them.

This study involved a variety of models both simulation and static and represented cost, product mix, and operational concerns. The output shown here is a snapshot of some of the results from the DES models.

As shown in Figure 5, factory scenarios are distinguished one from another according to their layout (flow versus functional) and the WIP machine assignment strategy used (tied versus untied). Notice there is no re-entrancy in the assembly process flows (in contrast to the highly re-entrant flows of wafer fabs).

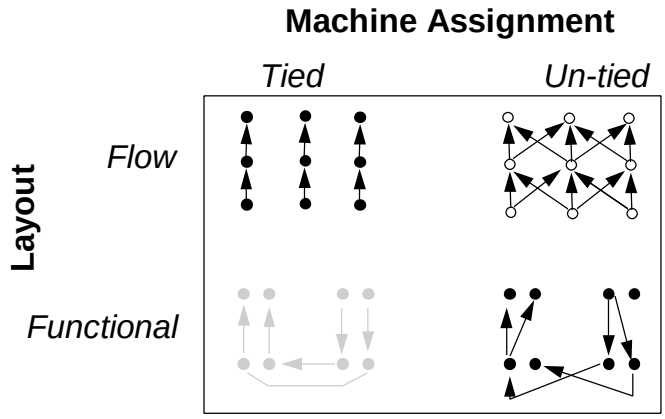

Figure 5: A conceptual view of the Factory Scenario Matrix showing the different conceptual cases considered (the functional-tied case was not considered)

A wide variety of experiments were performed assessing and ranking various WIP management strategies and understanding the impact of equipment reliability on overall factory performance. The "flow" test factory is laid out as four distinct manufacturing lines, with machines laid out sequentially in the order in which steps are performed in the manufacturing process.

In all models, a lot arriving at a step will not begin processing unless a "kanban space" at the output of the step is vacant. All models contain the exact same equipment set and have the same total number of kanban spaces available, so the maximum possible factory WIP is the same in every case. A common batching policy intended to maximize factory output while minimizing TPT is incorporated into all models where batched operations exist.

Figure 6 is an example of some of the output from the models. It shows the throughput time (relative to a base value) as a function of output for a variety of layout and operational scenarios. (For further information regarding the background of the ToC or Theory of Constraints operations point, see Karl Kempf's paper "Improving Throughput Across the Factory Life-Cycle" also in this issue of the *Intel Technology Journal*.)

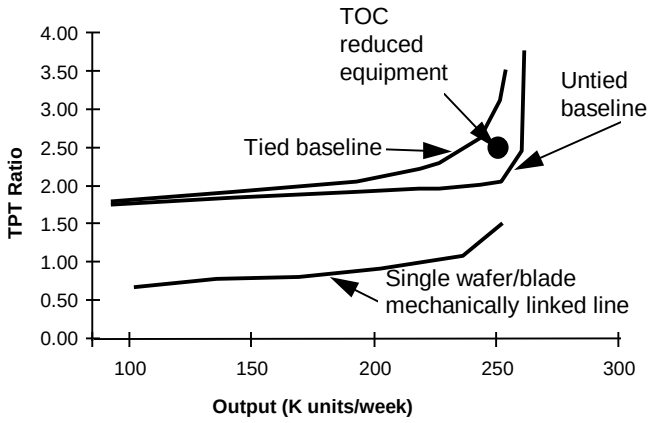

Figure 6: A comparison of baseline cases and ToC and mechanically linked-line scenarios

An examination of Figure 6 might suggest that one should operate the floor as a single blade mechanically linked line (continuous flow manufacturing). It turns out, however, that this option was much more expensive and resulted in a poor overall strategy because it did not have the required flexibility to handle the variety of expected product mix.

Other key learnings included the following:

- WIP management policy is more important than layout for this class of operation.
- Material transport, while affecting TPT and output, is not a primary modulator of performance unless it becomes very protracted.
- For the conditions studied, the functional approach outperforms the tied-line approach even when the number of equipment failures is cut in half. This suggests that simple improvements in equipment would not substantively alter the model-based recommendations.

Operation-Robot WIP Flow Interaction

A suite of three models was developed to allow assessment of WIP flow, TPT, robotic link performance and capacity issues in a key factory operation. The model quantified the impact of changes to the lot cascading ratio (a form of batching) to be about 20%. The model also highlights changes in PM and logic strategy resulting in a 10% capacity gain. These models were used to propose modifications to existing control software algorithms used by the robotics vendor. The models were used to quantify gains from eliminating specific steps and other conservative monitoring practices. This model showed that the suggested policy of putting wafers "at-risk" by processing them while waiting for monitor wafers to be read needlessly exposed the factory to line-

yield risk without any compensating productivity gains. Model use allowed us to propose novel methods for managing equipment setups.

Implant Area WIP Management Simulation

A DES model for an ion implantation processing area was based on the ability of a variety of diffusion and implant engineers, supervisors, and operators to determine the best area-WIP management rules. The model allowed implant staff to understand the impact of various WIP policies on productivity. It has highlighted a future equipment shortfall that no "creative" WIP management policy would alleviate. This model was also used as a decision-support tool to guide selection of current and future WIP management rules, and finally, it was used to guide prioritization of floor activities in the face of serious WIP surges caused by volume-ramping interactions.

Capacity Analysis Based on the Theory of Constraints—Simulation Studies

As mentioned earlier, our industry is challenged by the enormous capital costs of new factory equipment. Historically, Intel has estimated the amount of equipment needed based on a static and constant utilization-target-driven analysis (known as balanced line analysis). DES models were used to show that certain sets of equipment should be chosen as the factory constraints and that various other utilization targets should be used to assure a surplus of other equipment to optimize factory performance for a given cost, automation, and operations strategy. Based on the model findings, appropriate targets were selected for key equipment, with the expectation of a 15% reduction in TPT and a 50% reduction in variability of output for a 3% increase in capital cost over the balanced design [9].

Impact of Constraint Equipment Dedication Strategy on Fab Performance

As mentioned earlier, a unique feature of wafer-fab processing is the re-entrant nature of the material flow. For some of our fabrication processes, we require lot-level dedication of some manufacturing equipment at certain operations. This means that, when a lot returns to the operation, it must be post-processed on the identical piece of equipment on which it was pre-processed. An equivalent piece of equipment cannot be used. The intuitive expectation of the process development and operations staff was that significant factory capacity would be lost through the "interference" caused by these requirements. Additional factory space and very expensive equipment was placed into the purchasing cycle to compensate for these effects.

We built a detailed full factory model to assess the impact of this unusual constraint upon the re-entrant manufacturing line. It was found that there was no deleterious impact on factory performance, and we thoroughly understood why. While it is true that the amount of material in the queue at the dedicated steps increased and batching strategies and other resource issues needed to be optimized, we demonstrated that overall factory output and throughput time were not affected due to compensatory effects in other areas of the fab. In fact, we realized a savings of many millions of dollars by intercepting the purchase and installation of equipment for these fabrication lines.

It has often been remarked that hindsight is 20/20 and that we should have quickly realized this would be the case. One of the powerful benefits we derive from having complete and validated DES models is that we can develop correct intuition for cases like this where our own assessments fall short.

Factory Layout and Operation of the Single-Edge Connector Card Factory

Intel's Single-Edge Connector Card operation inherited much of its initial technology from our Systems (motherboard manufacturing) group. The Systems group had found very large gains when, years ago, they implemented mechanically linked continuous-flow manufacturing processes. It was assumed that as single-edge connector cards are like motherboards, their manufacturing operations should be designed in a similar fashion.

As part of our internal due diligence evaluation, we built complete DES and static models of the card operation that allowed us to study the performance of layout options similar to those shown above in Figure 5.

We found that the card operation would benefit greatly from not following in the footsteps of the motherboard line, and we found out why. The product mix of the card operation was significantly different from that of the motherboard line. Product differentiation in the market place drove constantly changing product requirements and time frames that could not effectively utilize all the capital equipment and labor that would be put in place to support a balanced and hard-linked line. We identified the optimal layout, batch sizes, buffering strategies (where needed), and operating policies. The verified models allowed us to thoroughly study the operations space and intercept factory build and fit-up plans, which resulted in a savings of many tens of millions of dollars in direct costs. Moreover, they gave us greater operational capacity and flexibility while still meeting our aggressive volume-ramp goals.

Operations Modeling Next Steps

As valuable as our model suites are and as useful as our teams of model-savvy engineers are, there is still much that remains to be done.

Quality and Voluminous Data

One of the largest hurdles inhibiting the development and application of DES models within a factory is the difficulty in obtaining detailed and accurate equipment and operations performance data in a timely fashion. These data must often be distribution based and not just a simple summary of means. The efforts of the many people within Intel working on common data definitions, common databases, and improved data gathering are of primary importance to the growing and widespread use of models.

Time Required to Create Models

There is a characteristic dissonance or tension that seems to always exist between model developers and those ultimately responsible for the tactical and strategic decisions. The decision makers want a model yesterday to help answer a problem they will first articulate tomorrow. The modelers, in turn, want months to develop validated models and gather certified data upon which multi-million dollar decisions can be confidently based. We are resolving this dissonance in two ways.

First, we are prebuilding a variety of models that, generically, we know will be of value. All future factory processes, when they leave our technology development site, will be transferred to high-volume manufacturing with a validated full-factory model already in place.

Second, as models always seem to take longer than expected to develop, we have an active program underway to ensure that past validated models can be rapidly put to new uses. We use standardized model languages. We are in the process of developing a common modeling framework and language interfaces to maximize our ability to plug-and-play various models and to model modules with each other. This framework must allow, at a minimum, the interfacing of models from different levels of abstraction. Our work to date has allowed us to solve problems in days and weeks that a year ago would have taken months to complete. More work is needed, however.

Advances in DES model performance and application areas, especially those associated with visualization, are expected to accelerate. These applications too must integrate into common data and language frameworks where possible. There are no industry standards to guide this work today.

The Speed of Model Experiment Execution

Model-based experimentation, while very fast compared to working directly with a factory, is still a bottleneck for various reasons.

First, it is difficult to determine the right level of abstraction to be used in a proposed model. This is largely an art now that is best done by our most experienced modeling engineers. Large DES models (due to detailed abstractions) take many hours of CPU time to run, require very large data sets, and may be complex to write. One would like to only use them when necessary and instead use fast analytic models, or models with less detail, whenever possible. We have studied this problem and for the time being have opted to model all the detail we can afford. We set priorities as to where detail is needed and we do all that the budget or time frame allows. Our experience when improving already highly effective wafer-fabs, running at full volume, has been that everything matters and interacts with everything else, thus requiring high-quality DES models. Linear flow assembly processes are not so sensitive.

The problem of long execution times can also be addressed by improving analytical models. In the past, queuing theory analytic solutions have not readily handled the re-entrant constraint of wafer fabs. Important work [10] in this area is now being tested in various proprietary settings.

Second, the synthesis approach of using the try-test-analyze and try again method of simulation model use is slow. Improved optimization techniques are required that take into account the classes of constraint problems that must be posed and at the same time can be integrated into our modeling tools.

Third, when one analyzes the results of a full-factory simulation, particularly when validating a model or studying dynamic changes, one may be analyzing an enormous volume of data. The analyst must have a thorough understanding of factory physics, be sensitive to subtle interactions, and be able to deal with mountains of numbers. Tools to assist in this process are needed.

The Chaotic Factory

Although we refer, tongue in cheek, to our jobs within manufacturing as challenging and chaotic, we may be more correct than we know. One often makes the assumptions that models and the factories they model are well behaved. That is, that small changes in input produce, generally, small changes in output.

Work by Beaumariage and Kempf [11] has started to carefully explore the notion of chaotic behavior within semiconductor fabs. It turns out that when factories

become very heavily loaded, performance may become unpredictable and strange in unexpected ways. DES models are excellent for identifying and studying this behavior.

Their study is founded on the mathematical notion of chaos theory, which says that a complex system can appear to randomly jump between a number of stable states with very little provocation. This system instability is a consequence of nature when large- and small-scale phenomena interact (for example, when multiple re-entrant flows interact with local equipment and operator issues).

The researchers began to suspect that formal chaos was present in fabs when they observed in simulation models that optimum schedules for heavily loaded Intel fabs changed dramatically with only slight changes in input. They developed specialized small DES models that capture key fab behavior. With these models, the team was able to observe and study transitions between multiple stable states, each with its own performance profile. They found, for example, that changing the order of one lot in a queue was sufficient to move the model to a state where average TPT was increased by up to 50%. In other cases, they demonstrated transitions between states with similar average TPT performance, but with widely different variability week to week. At least part of the chaotic nature seems to be aggravated by very high equipment utilization and by issues surrounding re-entrant flow. It is not yet clear at what level these findings are relevant in a real operating factory.

Labor

Most model environments treat labor as though it were a machine or a machine resource requirement like a jig or fixture. This is a large shortcoming when modeling loaded factories in detail. People are not automatons. They plan, they create, they locally optimize, they preempt, and they "glue" together situations that might otherwise reduce a factory's performance. An increased understanding of the human role and the degree to which it affects a factory's bottom line is needed. Kempf [12] has done some recent work in this area.

Operations Modeling Cautions

A few words of caution are in order for those who are establishing modeling centers within their companies. While it is true that having validated full-factory models and other models available for use is of tremendous value, it is not without risk. These risks include:

- *Incorrect understanding of the nature of the problem*: It is very easy for us "carpet dwellers" (as we affectionately refer to ourselves) to easily lose touch with what is really happening on the factory floors of the operations we seek to model. We find it essential to have floor operations' people intimately involved in all our projects and to ensure that our modeling engineers spend time on the factory floor. At the same time, we find it is very easy for factory and other personnel to not understand the full interactive physics of the factory, to create models that are simplistic, and to interpret the output of correct models in an incorrect fashion.
- *Non-validated models*: It is very tough to create a validated model. A validated model means that there is a preponderance of evidence to suggest that the model actually behaves like the real world in the area of interest. We have been involved in the "retreading" of a number of models from other companies or organizations due to validation problems with their work. Validation practices must be carefully thought out and religiously adhered to.
- *Controlling expectations and resourcing of projects*: To many people, good models may be magic. There is an almost universal lack of understanding of how a model can be correctly developed and used. Its use depends on the assumptions with which it was constructed, on the data available, and on the skill of the analyst applying the tool. Its development may take several people many months of effort. The temptation is always present to cut corners and deliver results sooner. This is almost always a mistake. To provide a model-based answer that is significantly in error is to perhaps fatally undermine the organization's modeling efforts. We find it very helpful to align our modeling efforts with the multi-year strategic roadmaps of our customers so that when their need arises we are already ready.
- *Starting with a focus on tools rather than on understanding the problem at hand and its business implications*: We use a formal contracting process with our internal customers to define clearly the business need, expected value, data and other information required from the customer, timelines, and deliverables. This agreement must be signed by the manager of the responsible customer organization and by the manager of the central modeling group.

Conclusion

Modern semiconductor factories are far too complex and costly to be optimized without the use of validated models. Models can be effectively applied to any level,

from the enterprise level down to the level of a specific robot within a piece of process equipment. Discrete-event simulation models are able to represent the richness of an operating factory and help provide insight into the dynamic response and optimal operation of the factory floor. Intel's use of manufacturing operations models saves millions of dollars in direct and avoided costs each year.

Acknowledgments

I am grateful for the excellent work of the many model developers, engineers, and manufacturing personnel whose work is referenced here. It is a pleasure to report on their vision and contribution. In particular the work of Sean P. Cunningham, Lance I Solomon, Edward J. Yellig, Erik A. Stewart, Karl G. Kempf, Gregory J. Mazenko, Rich Polliak, Mark F. Schaeffer and his Fab 11 team, KC Yoon and his Penang team, Bill Hendricks and his Chandler team, is gratefully acknowledged.

References

[1] Dickens, C., *A Tale of Two Cities*. Book the First, Chapter 1, line 1 (Gutenberg e-text January 1994, used as source)

[2] Fuhs, C., *SEMI-Advanced Semiconductor Manufacturing Conference*, Boston, September 10-12, 1997.

[3] Hopp, W., and Spearman, M, *Factory Physics*. Irwin (a Times Mirror Company), 1996, p. 1.

[4] Modified from: Cunningham, S. P., Internal communication.

[5] Miller, R., *Manufacturing Simulation*, The Fairmont Press, Inc., 1990.

[6] Poliak, R., Internal communication.

[7] Hilton, C., Mazenko, G., Solomon, L., and Kempf, K. *Assembly Floor Layout and Operation: Quantifying the Differences*, ISSM 1996, Tokyo.

[8] Deneb Robotics, Inc., 3285 Lapeeer Rd. West, PO Box 214687, Auburn Hills, MI 48321-4687 USA, 810-377-6900.

[9] Srivatsan, V. et. al., Internal communication.

[10] Kumar, P.R., "On the guaranteed throughput and efficiency of closed re-entrant lines" to appear in *Queueing Systems: Theory and Applications*. Paper is at http://black.csl.uiuc.edu/~prkumar/postscript_files.html

[11] Beaumariage, T., and Kempf, K., "The Nature and Origin of Chaos in Manufacturing Systems," *5th Annual !EEE/SEMI Advanced Semiconductor Manufacturing Conference and Workshop*, 1994, p. 169, Cambridge, MA, USA.

[12] Kempf, K., and Spier, J., "Simulation of Emergent Behaviour in Manufacturing Systems," *6th Annual IEEE/SEMI Advanced Semiconductor Manufacturing Conference and Workshop*, 1995, p. 90, Cambridge, MA, USA.

Author's Biography

Courtland M. Hilton joined Intel in 1981 after receiving his Ph.D. in chemical engineering from the University of Illinois at Urbana-Champaign and prior degrees from Brigham Young University. Court has worked in a variety of positions from research and development, factory engineering management, knowledge applications and strategic information technologies to his current position as manager of Intel's Manufacturing Operations Modeling and Simulation Group. He is active in the SRC (Semiconductor Research Corporation) and is a popular and invited guest lecturer on many topics related to manufacturing science. He has served as an adjunct professor of statistics and is a recipient of Intel's Individual Achievement Award.
His email address is courtland.m.hilton@intel.com.

Planning for the 300mm Transition

Daniel Seligson, Technology and Manufacturing Engineering (TME), Intel Corp.

Index words: 300mm, manufacturing, cost, automation, standards, consortium

Abstract

Beginning in 1993 a small group of people at Intel began thinking seriously about a transition from 200mm manufacturing to the next wafer size. By early 1994, the industry reached consensus that the right size was 300mm. Late that year, we formed a cross-disciplinary team whose responsibilities included defining the 300mm goals for each of six functional domains: equipment, automation, factory and facilities design, EHS, manufacturing operations, and materials. In addition, we built a cost model that helped inform the above process while providing indicators of success that cut across all the domains.

The defined goals were widely disseminated in the industry through various channels including supplier management organizations and consortia. The equipment selection process was augmented to include the new requirements for tools, and the processes were extended to include all tools needed for a factory, not merely the most expensive and technically sophisticated. Task forces were established to answer urgent questions about lot size, carrier design, mini-environments, and any other issues that arose. In parallel, a new international consortium was built from the existing SEMATECH infrastructure, although with more clear cut goals and achievable objectives than SEMATECH had. This new organization, called I300I, became a critical part of the industry consensus building and supplier management. While we have gone through major industry transitions between program start and now, in areas such as market segmentation, transition timing, and cost focus, the goals developed in 1995 remain sound. Today, selection teams are in place for all equipment, from materials handling to clean parts, many of them selecting tools for new semiconductor processes never implemented at Intel. The teams are synchronized to produce a first set of tools in January 2000 and high volume 0.13um manufacturing capability in the second half of 2002. The productivity, as measured by reduction in die cost, is expected to exceed our targets of 30%. For comparison, only about a 15% productivity increase was achieved in moving from a 150mm to a 200mm wafer size.

Introduction

The goal of this paper is to provide a comprehensive summary of the critical issues and key events of the transition to 300mm. We begin with the decision to move to a new wafer size and how we arrived at 300mm. We then discuss the consensus that this would be an industry transition, rather than one led by a single company. The process by which Intel came up with the requirements we tried to drive in the industry is reviewed next, followed by a brief look at the cost model and productivity. Automation, in particular automated materials handling, received a lot of focus in our planning. Its importance and implications are explained in the next section, together with an explanation of a major improvement in factory design. Other equipment, safety, and factory requirements are then mentioned, where we also provide a one-page summary of all requirements. Getting the message distributed within Intel and the industry was a huge effort, so a special section is devoted to how that was done. The two consortia, I300I and SELETE, played large roles that cut across many aspects of planning, so we also review their contributions. We then discuss the many improvements and the entirely new business processes that have been put in place to help manage the transition as a whole. In the final two sections, we cover the high-level review processes Intel has used to make overall program decisions, and the outlook for the future from the vantage point of Q4 1998 is briefly discussed.

Selecting the Next Wafer Size

Intel began 200mm manufacturing in 1993, following a two-year development effort. The larger wafer size gave us nearly twice as many die for every wafer moved, but the growth in the microprocessor business and the growth in die size led us to conclude that before the end of the decade we would need to be adding factories at the rate of two per year. The complexities of construction, site selection, staffing and management development required to support that growth suggested that it was time to begin thinking seriously about the next wafer size.

Many of the costs of manufacturing are proportional to the number of wafers moved, and not to the area processed.

So, size does matter. Initially, Intel favored 400mm and Applied Materials, the world's largest semiconductor equipment manufacturer, put forward arguments for 350mm. But the silicon wafer manufacturers told us that if we were contemplating a wafer size change within a decade, the largest wafer we could have was 300mm.

Their arguments were based on the fact that the length of the boule should increase at least in proportion to the new wafer size, and its weight therefore by the cube of the dimension. The starting material for a 300mm boule will be between 300kg and 450kg. New inventions would be required to cost effectively manufacture wafers larger than 300mm. The correctness of this view can be seen in the existence, schedule and goals of the Japanese Super Silicon Crystal Institute Corp, a partnership of government and industry that focuses on developing wafers of 400mm and larger by 2004.

Process equipment was predicted to have no particular issues with any of these wafer sizes, although initially there were fears that the gravitational sag of hot wafers in a furnace would induce slip. The flat panel display industry was at that time using substrates in which a 350mm circle could be inscribed easily, so 300mm looked well within reach technically.

An Industry Transition

The transition from 100mm to 150mm in 1983 and 1984 was led by Intel's groups in Albuquerque, New Mexico. There is no historical data on its cost effectiveness. The transition to 200mm was led by IBM in Burlington, Vermont. First tools were delivered in January 1986, and 1 Mbit DRAMs were first qualified for production in early 1998. When Intel made this transition with production beginning in 1993, die costs were reduced by 10% to 15% when comparing new 200mm and new 150mm fabs. But when making the more realistic comparison of a new 200mm fab to a 150mm factory upgraded to meet new technology requirements, no die cost improvement was achieved. IBM may have achieved a greater reduction because they converted from 125mm wafers.[1] Both transitions were unpleasant experiences for the lead company in that it had to bear the burden of development costs, manufacturing delays, and poor equipment performance, all at little or no cost benefit.

These unfortunate precedents led us to conclude that no single company was smart enough or large enough to do a wafer size conversion by itself, and that the 300mm transition should be an industry one. The hoped for benefits were common performance objectives, shared learning, cost sharing, and more efficient and accelerated development facilitated by use of widely accepted standards.

Standards were expected to be important in such diverse areas as wafers and wafer carriers, data and material transfer protocols, safety, and others. Historically, standards had been defined *post facto* on the basis of successful implementations. Early implementers were almost guaranteed to be outside the standards. For this transition, the industry believed it would be better to define the standards at the outset and ensure that everyone adhere to them. SEMI, the international standards organization, was thus destined to play a key role.

Shared learning meant that we would do as much pre-competitive work as possible[2], particularly in the testing of equipment and modular components. The results of these tests would be fed back to the supplier, along with a roadmap for improvement. They would also be made available to the members of the industry group performing the testing, although at that time the group was not yet defined. The initial vision was of an Underwriter's Lab™, providing a stamp of approval instead of a *caveat emptor*.

Cost sharing really meant that the equipment suppliers would pay for the equipment development, and a manufacturers' consortium would pay for its testing. During the two previous transitions, when Intel and IBM paved the way, the suppliers were relatively small and weak. By the mid-90s, several of the suppliers had annual revenues exceeding $1B, with R&D budgets capable of supporting development of a new generation of tools. The increased health of the supplier base can be attributed to prolonged growth in the semiconductor industry, consolidation in the equipment industry, and, in the US at least, the attentions of SEMATECH. Such a model could not have worked for the 150mm and 200mm transitions.

Finally, we recognized that having common performance objectives for tool performance would be a critical element of any industry transition. One of the lessons of SEMATECH was that everyone's secrets were the same, but they resided

[1] Die cost reductions are driven largely by the ratio of the number of die on the final and initial wafer sizes. For the three transitions 125->200, 150->200 and 200->300, these ratios are approximately 2.7, 1.9 and 2.4 respectively. The exact ratios depend on company-specific product issues.

[2] Pre-competitive work was not possible during the previous wafer size transitions. Not until the semiconductor crises of the mid 1980s were laws passed to allow the pre-competitive work such as that done at SEMATECH.

separately in each company. This compartmentalization of knowledge increased the workload on the supplier and decreased efficiency. Working through the difficult early years of SEMATECH, the US manufacturers learned to give direction to the supplier base without compromising their individual intellectual property. This greatly simplified their workload and guaranteed, on average at least, a higher quality product. Since the previous wafer size conversion had shown modest paper cost benefit and less real cost benefit, a higher quality product was badly needed.

An industry transition also meant that no one company would be first, or would need to be first. Nevertheless, by late 1994, Motorola had established itself as the industry leader, although they participated in the various standards' bodies and consortia. Their aggressive schedule and forcefully stated opinions helped accelerate the pace of decision making on important issues. Other companies chose to say that they wouldn't be first. Later, when Motorola's program was put on hold, enthusiasm still remained high. The pundits concluded that no company wanted to build the first 300mm fab, but neither did any company want to build the last 200mm fab.

Developing Intel's Requirements

In late 1994, an *ad hoc* 300mm discussion group, involving Process Equipment Development (PED) and California Technology and Manufacturing (CTM), realized that because 300mm fabs were still far off in our future, it was an opportune moment to begin developing a vision of a 300mm world. We invited representatives from all of Intel's stakeholding groups to the discussion. Since we had no authority to command, those who joined us were only the 'can do' types who wanted to be there, so progress was swift. By early 1995, we had built a group that covered major problem areas and so took on the task of writing a handbook defining our requirements in six different functional domains: equipment, automation, factory and facilities design, EHS, manufacturing operations, and materials. By mid-year, we had named ourselves the Cross Functional Working Group (CFWG) laying ourselves crosswise on a set of committees (SCS) whose job it was to select new tools and manage roadmaps for the evolution of process equipment, automation, factory design, chemicals and their use, manufacturing operations, and so on.

The requirements for 300mm automated materials handling, for example, would necessarily add requirements to the process equipment whose roadmaps were owned by an equipment SCS. This necessitated many rounds of negotiation among the various SCSs, complicated by the fact that the SCSs had neither the time nor inclination to listen carefully to arcane discussions about a wafer size that existed on no Intel plans. Nevertheless, over the course of 1995, each domain developed its requirements' package, reviewing and refining the material at the CFWG, an audience increasingly attentive to the full spectrum of issues relevant to a 300mm transition.

Several problems emerged whose solution was outside the range of any one existing group. To solve these problems, we established task forces. Questions of lot size, lot buffering, lot carrier design, reticle carrier design, assembly components, and mini-environments were answered in this way. When relevant, our task forces would work closely with industry standards' organizations too. In the case of lot size, industry opinion ranged from 1 to 50 wafers per lot in mid-1994. Single wafer processing experiments at Texas Instruments (TI) suggested that the low latency and short throughput times could be advantageous. Motorola advocated single wafer transport initially. Meanwhile, the experience of Intel and the large DRAM manufacturers was that larger lot sizes were favored.

Trying to separate the emotion from the facts, we did a numerical factory simulation, varying lot size from 13 to 50 wafers, assessing factory capacity and several cost drivers. Small lot size implies more lot moves per unit time. We set the lower bound of our study to 13 because we knew that smaller lots would overwhelm any realizable automated materials handling system (AMHS). The lot size team looked at issues including ergonomics, metrology, floor space, supplier capability, capital equipment costs, labor, and total wafer cost. They drew three important conclusions. First, ergonomic considerations would preclude regular manual handling of a lot with 13 wafers or more; therefore, extensive AMHS would be needed for any lot size. Second, total processed wafer cost decreased with increasing lot size, so larger lots were favored. Third, AMHS on suppliers' drawing boards would be unable to move 50 wafer lots; they were too heavy. The answer, therefore, was 25 wafers per lot, and that became the Intel position.

A memorable event took place in February 1995 at the SEMI Standards meeting in New Orleans. In a crowded stuffy room, representatives of Motorola, TI, Intel, the Japanese DRAM manufacturers, and several dozen equipment suppliers debated the merits of different lot sizes. Motorola had retreated from its vision of single wafer transfer and had begun promoting 13. Someone asked, "is there anyone here that favors 13 over 25?" In response, only TI would even entertain the discussion. Motorola was defeated on the standards front; they could marshal the industry by dangling the hope of purchase orders, but not on the strength of their technical arguments or their negotiation skills. Gradually, Motorola

began to withdraw further and further from the industry center choosing to go it alone, convinced that the consortia and standards activities would only slow them down. When the dust settled, the industry standard for lot size was not one size, but two, 13 and 25. This added unnecessary development costs in both dollars and time for the suppliers of process tools, loadports, carriers, and material handling systems, to name a few.

The lot size debate is not as clear as portrayed here. Semiconductor manufacturers with business models different from an Intel or DRAM manufacturer, who make large volumes of a small number of products, might arrive at a different lot size. These business model issues were not factored into our analysis. In addition, the impact of the assumptions we used, particularly those having to do with lot-based rather than wafer-based metrology sampling plans, was never examined. Nevertheless, there is wide consensus today that 25 is preferred, although it introduces difficulties for companies manufacturing a large number of different short-run products.

The methodology used to make the lot size decision is typical of that employed for the other truly cross-functional decisions. It is not within the scope of this paper to include a detailed discussion of each, however.

The Cost Model[3]

As mentioned earlier, the move to 300mm was driven by the fact that it would provide an exit from the anticipated scenario of building two factories per year. Cost reduction was a factor, but more of an opportunity than an overriding force. To understand and then exploit the opportunity, the CFWG took on the task of developing a 300mm cost model. There was scant documentation of the cost targets for the 200mm transition and no documentation of performance against those targets, other than the overall null result. We believed that a more thorough modeling effort, followed by aggressive goal setting based on model parameters and detailed management of progress towards those goals, would yield a more felicitous outcome.

Starting from a 0.35um 200mm wafer cost model, which was all that was available at the time, we examined the major line items. These line items included, among others, capital depreciation, direct and indirect materials, utilities consumed, factory infrastructure, labor, and site overhead. We challenged ourselves to imagine a 0.35um 300mm cost model and to discover the mechanisms that underlie the scaling of each line item as we move from the smaller to the larger wafer size. By operating at the same technology generation, we were able to separate the wafer size issues from the process technology issues. For each line item we had therefore not merely a scaling number, but a concept.

This turned out to be quite powerful, for instance in modeling labor costs. Received wisdom had it that labor content per wafer increased 20% when we had increased wafer size in the past. So, we put into our model an equation looking something like this:

$$\text{labor}(300) = 12 * \text{labor}(200) \quad [1]$$

Over time we realized that this factor of 1.2 was a parameter,

$$\text{Relative Labor} = 1.2 \quad [2]$$

that we could control, not necessarily something predetermined. Achieving less than 1.2 would require that our operations' groups put plans in place to make it happen, possibly in conjunction with an AMHS group. Operations absorbed this and asserted that the right goal was parity or better as in,

$$\text{Relative Labor} \leq 1.0 \quad [3]$$

In this way, working with stakeholders, we set goals for all line items and for the groups that had a stake in them. The groups have gradually accepted more aggressive goals as they see the need, the opportunity, and the way.

Table 1 is an example of the methodology applied to a 0.25um DRAM process[4]. It illustrates that the scaling is substantially different for different line items. It is typical of what we see in the full-blown model Intel uses for planning purposes.

Line Item	**200mm**	**Scaling Factor**	**Test wafers**	**45**	
300mm					
	$	%		$	%
Depreciation	793	41	1.50	1189	35
Labor	232	12	1.00	232	7
Maintenance	155	8	1.50	232	7
Consumables:					
Direct materials	90	5	4.50	405	12

[3]For a more complete analysis, *see Semiconductor International*, January 1998, Daniel Seligson, "The Economics of 300mm Processing."

[4] The underlying 200mm DRAM cost components are drawn from Jack Saltich's paper in the proceedings of ISSM '94. This table was published in Semiconductor International in January 1998.

Test Wafers	45	2	4.50	203	6
Indirect materials	445	23	2.00	890	26
Other	174	9	1.30	226	7
Totals	1934	100		3378	100
Equivalent				1407	73

Table 1: 0.25um DRAM cost model for 200mm and 300mm wafers

Finished wafer cost in a new factory is dominated by equipment capital cost. Meanwhile, the desire to minimize the number of new factories was the original driving force behind the move to 300mm wafers. For this reason, two parameters that have received special attention have been the Relative Capital Cost and the Relative Footprint of the toolset. With X referring to either Capital Cost or Footprint, these parameters are defined as

$$\text{Relative } X \equiv \frac{X(300)}{X(200)} * \frac{OutputCapacity(200)}{OutputCapacity(300)} \quad [4]$$

Practically speaking, Relative Capital Cost is the ratio of the capital costs required to build a 300mm and a 200mm factory, each having the same number of wafer starts per unit time. Similarly, Relative Footprint is the ratio of the two factory areas. While defined for the factory in aggregate, these parameters can be measured for any individual tool. A customer can easily compare product offerings from multiple suppliers if consistent 200mm normalizations are used. The model parameters are easily visualized, as are their knobs.

One of the most powerful messages of the 300mm transition emerged from the realization that the value of these two parameters would determine whether the transition met its twin goals of increasing factory and capital productivity. The die cost model could thus be used to identify the parameter values required for success. The linear die cost model, which predicts costs given parameter values as input, could be inverted to produce parameter values given cost as input. Contours of constant die cost are lines in the space of Relative Capital Cost and Relative Footprint. Our senior management challenged us to deliver more than 30% die cost reduction, more than twice the reduction that had been expected in the 200mm transition. The inverted model told us that the Relative Capital Cost and Relative Footprint had to be below the red line in Figure 1. (In this figure, each point represents a different kind of process equipment. The area of plotted points is proportional to the product of individual tool cost and the number of tools required for a high-volume factory.)

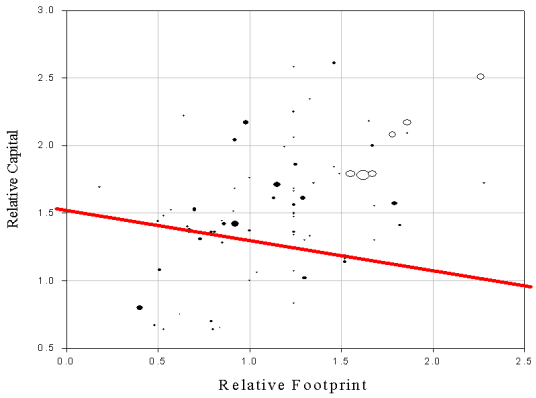

Figure 1: Productivity scaling factors for tools in a 0.18um logic process flow (data from early 1997)

This in turn gave us a simple tool to communicate with suppliers and the industry. On the whole, the new toolset needed to be below the line or we wouldn't get the sought after return, and we wouldn't make the transition. The supplier could quickly determine, by using equation 4, where they stood with respect to the line. The line became a high-level design target which provided guidance on the tradeoff between cost and footprint. Further simplifying the message for greater impact, we focused on a single point on the line, Relative Capital Cost = 1.3 and Relative Footprint = 1.0. Our vision of the future consisted of a 300mm factory that was the same size as our 200mm factories, producing the same number of wafers per week, and requiring a capital investment no more than 30% larger, as shown in Figure 2.

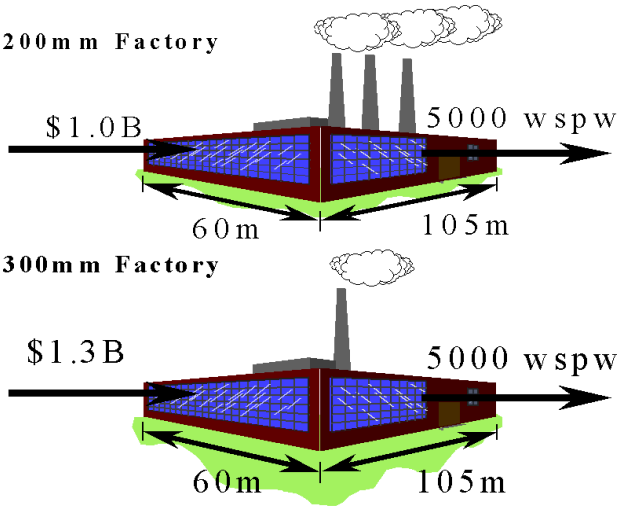

Figure 2: The macro view of the 300mm vision in which 200mm and 300mm factories are compared

When estimating die cost, two different scenarios are considered. In the first, you compare new different wafer size factories. This is the Greenfield Scenario, and our target for this is a 30% die cost reduction, obtained at a Relative Capital Cost of 1.3 and Relative Footprint of 1.0. In practice, the Reuse Scenario is more realistic. In this, you compare a new 300mm factory with an existing 200mm factory, one whose equipment set has been upgraded to meet the new technology requirements. Upgrading one of our 0.18 or 0.25um production lines to 0.13um technology results in a line whose capital basis is about half that of a new 200mm line. Figure 3 shows die cost vs. Relative Capital Cost for both the Greenfield and the Reuse Scenarios. It illustrates the point that the Relative Capital cost must be approximately 1.3 in order for the transition to achieve any real return. It also explains why the 200mm transition, with its 15% die cost improvement for the Greenfield Scenario, returned no benefit in practice.

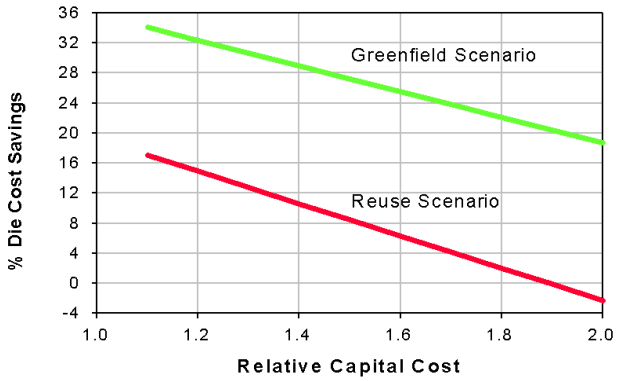

Figure 3: Die cost reduction for two conversion scenarios

Cost is important, but other measures are important too. What we are really talking about in this transition is increased productivity, where productivity is defined as follows:

$$\text{Productivity} \equiv \left(\frac{\text{something}}{\text{die out}}\right)^{-1} \qquad [5]$$

Capital productivity increases when the Relative Capital Cost decreases. Factory space productivity increases when the Relative Footprint decreases. Some of the other parameters we have measured and tracked include the productivity of chemicals, natural resources, utilities, and labor, as mentioned earlier. Just as we set values for the capital and footprint productivity and put plans in place to achieve them, e.g., through our supplier management organizations, so also have we done for these other measures of productivity. The 300mm transition is a watershed of productivity improvements for the industry. To date, all of the productivity targets we set in 1995 appear realizable or very nearly so, suggesting perhaps that we didn't reach far enough.

Over the years of 300mm planning, responsibility for managing the cost model has moved from engineering to finance, which is where it resides today. Its evolution has depended on maintaining a close relationship with engineering.

AMHS and a New Factory Design

In 200mm, we saw the first widespread use of AMHS. There were industry standards for such simple but important features as the height at which wafer carriers, or pods, rest on tools and the orientation of those pods with respect to the tools. But these standards were not implemented uniformly, making intrabay AMHS solutions custom, and therefore expensive, in every case.

Based on favorable outcomes with 200mm intrabay delivery to diffusion furnaces, with improvements in labor and capital productivity, we worked hard to define a set of AMHS standards that would facilitate 100% intrabay delivery. This was also desirable from an ergonomics point of view, as discovered in the lot size analysis[5]. It was important to get the industry to agree on what was required and to communicate those requirements to the suppliers before the equipment arrived. Such a set of standards, focussing on the tool loadports, would decouple the AMHS from the equipment, but it would also be essential that the standards were implemented by all suppliers. One example of how AMHS standards lower cost is that they enable a mechanically simple 1-axis robotic transfer from AMHS to tool to be used, rather than the 6-axis robots required in our present day 200mm implementations. For nearly two years after the standards were defined, managers at Intel would ask, "How many of the tools will arrive with the standards in place?" The answer, "They're requirements, so 100%," would be followed by challenges based on prior history of meeting such commitments, particularly in the area of automation. Today we all accept that the answer is 100%. That result is the product of extensive internal and external education forums and supplier management programs. Figure 4 illustrates the main points of the AMHS and loadport interface standards.

[5] We struggled to find the right justification for 100% intrabay delivery. A return on investment analysis depended on too many unknowns. Ultimately, Manufacturing Operations' policy of striving towards an incident and injury free workplace provided the justification we have come to accept.

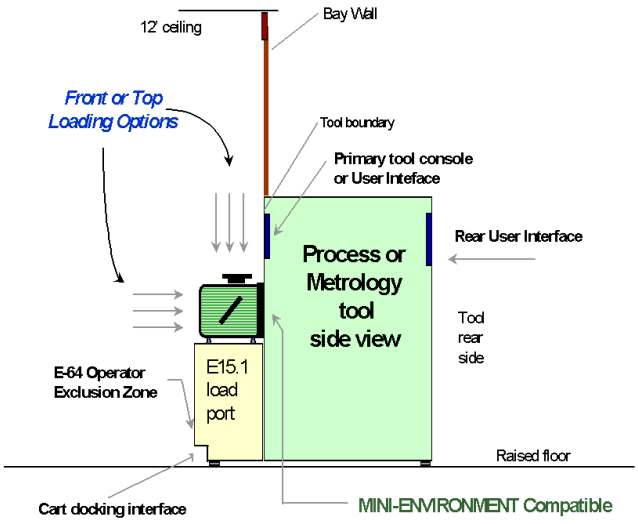

Figure 4: Features of the AMHS and Loadport interface standards (picture courtesy of D. Pillai)

Independent of 300mm planning, Intel has established an ongoing benchmarking program whose objective is to gather information about the relative performance of Intel against other semiconductor manufacturers. In the mid-1990s, while 300mm planning was blooming, this group learned that our fab construction costs were quite high compared to the best manufacturers. A task force was established to develop a new factory design for 300mm, setting cost targets 30% lower than our most recent 200mm factory. Prior to this goal being set, the initial value for the Relative Factory Costs in the cost model was 1.08, much worse than the task force's target of 0.70. The value 1.08 was based on historical precedent. This serves to illustrate how costs had spiraled upwards in the past and how 300mm successfully reversed the trend.

Similar to the CFWG, the task force drew on all stakeholders, with a surprisingly large amount of input from AMHS. The result (see Figure 5) was a unique, and for Intel, radical design. It disposed of many features that had previously been considered essential. Further, it included novel solutions to problems brought about by 300mm, in particular problems related to the fact that lot storage space requirements would increase significantly. The final cost, although still only an estimation, was 0.62, beating the task force target.

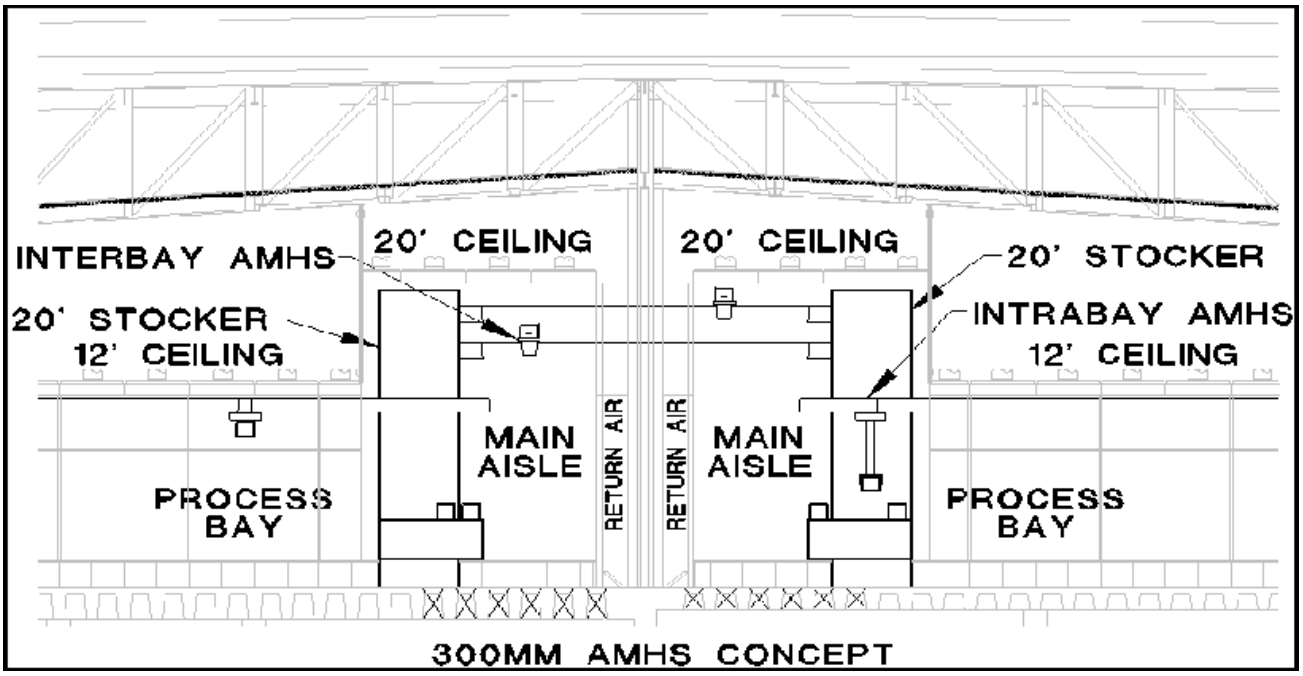

Figure 5: Cross section (transverse to the main aisle) of the new 300mm factory concept. Note the tall center section providing efficient storage of WIP. Figure courtesy of Noel Acker.

Other Equipment Performance Requirements

Chemical usage and environmental stress had become an increasingly large issue, accompanied often by negative publicity. Analyses showed that approximately 30% of our factory building costs were tied up in facilities that managed the ebb and flow of chemicals and utilities. We extended the notion of productivity to these items, establishing targets for the use of hazardous air pollutants, electrical power, scrubbed exhaust, and others, all measured relative to 200mm and normalized to capacity. For some of these, we didn't know whether the targets were achievable, and the measured data is not yet in, but the projections are that we will come close, resulting in an overall reduction of more than 50% per die.

In order to reduce the costs associated with the installation and qualification of tools, we developed a set of guidelines for standard practices, as well as setting relative targets calling for a 30% reduction in time and cost. We have since been working closely with suppliers to help them understand the guidelines, forecast their costs, and address issues when the forecasts fail to meet the goals.

The handbook of these and other requirements formed a thick presentation that took more than four hours to deliver when first rolled out to some suppliers in November 1995. This was useful but tedious. We realized that we needed a one-page document understood by everyone and capable of being carried around in an executive's shirt pocket. Table 2 is the one-page summary containing all the key requirements from all the participating groups. We identify the relevant industry specs, and we include a column identifying the I300I position on each requirement. (I300I is an international consortium that is explained later.) The I300I column emphasizes the fact that our requirements are more than Intel's requirements. Ultimately, they became requirements for the entire industry.

300mm Equipment Performance Requirements

Domain	Topic	Specification	References	i300i Position
EHS	Safety and Ergonomics	100% Compliance	SEMI S2-93A and S8-95	same
	EU Machinery, EMI and Low Voltage Directives	100% Compliance	EUVCE, E33-94	same
	Relative Recordable Injury Rate	<1.0		
	Relative Hazardous Air Pollutants Emissions per wspw	< or =0.5		same
	Relative Perfluorocarbon Emissions per wspw	< or =0.5		same
	Relative Volatile Organic Compounds Emissions per wspw	< or =0.5		same
	Relative Emissions and Utilities Consumption per wspw (e.g., water, process electrical, scrubbed exhaust, bulk gases)	< =1.0		same[1]
Equipment	Relative Factory Footprint per wspw	<1.0		same
	Relative Capital Cost per wspw	<1.3		same
	Capital Equipment Availability	>90%		same[2]
	Tool Install/Qual Cost as % of tool cost	<6%		same
	Relative Tool Installation Duration	<0.7		same
	Relative Tool Qualification Duration	<0.7		same
	Relative Spares and Maintenance $/wspw	<1.0		same
	Relative Monitor Wafer Usage per wspw	<0.25		same
	Edge Exclusion	3mm		same
	Electrical Voltage Drop-out Immunity	100% Compliance	IEEE-446 (CBEMA) Curve	same
	Mini-Environment Compatibility	100% Compliance	SEMI E44, E62 (FIMS)	same[3]
Automation	Load/Unload Port Interface	100% Compliance	SEMI E15.1-0697	same
	Integrated Box Opener/Loader	100% Compliance	SEMI E63 (BOLTS)	same[4]
	Parallel I/O Interface for Automatic Load/Unload	100% Compliance	SEMI E23-96	same
	Cart Docking Interface	100% Compliance	SEMI E64	same[5]
	Lot Size	25 wafers	SEMI E1.9-0697	13/25[6]
	Lot Buffering	2 Unicassette Ports minimum with capability of cascading. (Note: High runrate tools will need special attention.)		same
	Carrier Architecture	Horizontal transport, Front loading, Cassetteless, 10mm pitch, Kinematic coupling	SEMI E62, E57-1296, E47.1-0697, E1.9-0697	same[7]
	Embedded Controller	SECS II, GEM, HSMS	SEMI E5-96, E30-95, E37-95, E37.1-96	same[8]
	Network Connectivity	RS-232 & Ethernet with TCP/IP		

Table 2: 300mm equipment performance requirements

Communicating the Requirements

The SCS-ratified CFWG requirements were communicated through an extensive infrastructure of existing supplier management channels, primarily through the department that is now called Capital Equipment Development (CED). A small group of 300mm content experts teamed with engineers and managers responsible for the performance of individual suppliers to deliver the 4-hour requirements package to each of more than a dozen suppliers by the middle of 1996. These meetings served to educate our own teams as well, since 300mm was new and generally of low priority compared to more pressing 200mm programs. We used these meetings to understand the suppliers' 300mm program status and their reactions to our requirements as well. Management of supplier performance to the requirements was, at least in principle, done not by the 300mm content experts, but by the individual supplier team. We produced a video to facilitate delivery of the handbook content both within Intel and to the supplier base. We have also held topic-specific meetings with a large community of suppliers, the purpose being to clarify particularly subtle aspects of the requirements.

The Consortia

In 1987, a group of American semiconductor manufacturers and the US government formed a consortium aimed at reversing the losses the US industry had suffered at the hands of the Japanese. The consortium was called SEMATECH (Semiconductor Manufacturing Technology), which celebrated its move into its Austin headquarters in November 1988. SEMATECH focussed on rebuilding the infrastructure of the American equipment suppliers. There is little doubt among the faithful that it was successful. Some argue, however, that the cost was excessive and that a set of conference rooms, a coffee machine, and a legal umbrella would have accomplished as much and would have saved us building a high-overhead factory.

By the mid-1990s, two developments were shaping some major changes within SEMATECH. Firstly, 300mm was identified as the next wafer size, so SEMATECH initiated a 300mm program. As mentioned at the outset, there was agreement amongst most players that 300mm should be different, and in particular that the players should cooperate as much as legally possible. Along those lines, in late 1994, a SEMATECH task force ran a series of meetings organized by process type. Suppliers and member company (MC) representatives were invited, and the agendas included identifying critical problems that needed resolution, prospective MC schedules, and performance specifications. In parallel, SEMATECH was preparing cost analyses, based largely on estimates from the suppliers of what they wanted to charge. A summary meeting was held at year's end, during which, among other things, suppliers and customers resolutely rejected the notion of bridge tools[6].

The second development was that the SEMATECH MCs and the US government decided to end their partnership. This naturally led to the idea that a SEMATECH 300mm program should have international membership. Combining that notion with the desire to have a consistent testing methodology, the international 300mm initiative or I300I was spun off from SEMATECH. It would use SEMATECH infrastructure, but would be international and would secure most of its funding from its own MCs and not from SEMATECH resources. Ultimately, it grew to 13 MCs, 6 outside the US, but none from Japan. Dues were approximately $2M per MC per year. Initially, the scope of I300I was limited to delivering test results on a set of 0.25um 300mm tools by the end of 1997.

As SEMATECH began to deliver results in 1990, the Japanese bubble economy of the 1980s began to deflate. By the mid-1990s this was compounded, for our industry, by the amalgam of advancing semiconductor companies from Korea and Taiwan. Rather than join I300I, with its distinctly American flavor and slightly jingoistic aftertaste, the Japanese formed their own SEMATECH-like organization called SELETE, the Semiconductor Leading Edge Corporation. As SEMATECH did for the Americans, so was SELETE meant to do for the Japanese, i.e., give the (Japanese) MCs and (Japanese) suppliers a competitive advantage over their counterparts in other countries. Annual dues were approximately five times higher than those of I300I, but expected outcomes were greater: they included improvement and development of new tooling. SELETE was following the SEMATECH model while SEMATECH itself was moving away from it. Initially, SEMATECH had tried to rally the Americans with wartime fervor, citing the Manhattan Project and other heroic feats. SELETE perhaps was trying to recreate the heroic deeds of Japan's VLSI program in the late 1970s and early 1980s, which developed the 64k DRAM and which was responsible for vaulting the Japanese into a leadership position in semiconductors.

[6] Bridge tools are used across a wafer size transition; they had been commonplace at the 200mm transition, with the result that 200mm tools offered little economic advantage over 150mm tools. Bridge tools are making a bit of comeback today, but with a few exceptions, a single tool cannot meet our performance expectations at two different wafer sizes.

The stated purpose of I300I was to distribute, over the 13 MCs, the costs of evaluating tools, as well as to provide those evaluations in a consistent manner. To that end, an extensive Demonstration Test Methodology (DTM) was developed that provided MCs and suppliers with a common language to discuss and measure tools. The scope of the DTM, and the number of tools tested determined the I300I budget. Testing of tools was the prime deliverable, and the DTM has been applied to 60 demonstrations, of varying degrees of extensiveness, as of this writing.

There were two additional somewhat unexpected results that I300I delivered that have been very valuable. First, the schedule they set for the demonstrations became the *de facto* schedule for the industry conversion. Motorola was trying to move the industry forward with unilateral pronouncements, supplier meetings, and promises of purchase orders to come. Other manufacturers were announcing their own schedules with less fanfare. The result was that there was no industry schedule for this industry transition until I300I published a schedule for the 1997 demonstrations. This served to align the demands of the manufacturers and the readiness of the suppliers. Some have said that it introduced delays because leading customers backed off until after the demo results became available, but these original schedules were marketing schedules, not manufacturing schedules. In 1998, now that the time line has been pushed out twice, many suppliers are irate with I300I for publishing such unreliable schedules. However, I300I has merely repeated what its MCs have told it, and the forces driving the delay are powerful indeed, although analysis of those is beyond the scope of this paper.

The second additional result was that I300I has become the *de facto* standards body for 300mm worldwide. As part of its effort to develop the DTM, it had to identify *what* would be tested. For each different tool type, appropriate performance measures and target values, such as, etch rates, process uniformity, or capital costs per wafer per unit time, were developed[7]. In addition, I300I worked to develop a set of requirements (AMHS, safety, and other areas) common to all tools. A high degree of compliance to the common requirements, as determined by I300I, became the key to further testing. These requirements, agreed to by the 13 MCs, became the defining standard of 300mm. SEMI is the official international standards-setting organization of our industry. The SEMI balloting process was used to ratify what the customers had agreed upon at I300I.[8]

The fact that we had two consortia, I300I and SELETE, served to complicate matters because the I300I requirements could not truly be the industry requirements until they were negotiated with the Japanese. Additionally, SELETE explicitly stayed out of the game of setting standards, leaving that to another organization called J300. Teams from I300I and J300 began meeting by mid-1996, actively trying to avoid a built-in divergence in requirements. At SEMICON West in 1997, the two organizations announced and distributed their single set of requirements for the 300mm generation. These were called the Global Joint Guidance on 300mm Semiconductor Factories.[9] They are barely distinguishable from the set of requirements developed by our CFWG. At SEMICON Japan 1998, we expect a similar document to be published covering the requirements of 300mm assembly equipment.

The path from CFWG to Global Joint Guidance is our spec or standards pipeline. Following problem identification, we form a multidisciplinary group within Intel to develop a need-driven schedule to solve these problems. Some members of this group need to be well informed about and participating in relevant external activities and industry groups. If there were no such activity, they would initiate it. The group will meet regularly and review progress at the monthly CFWG meetings and if necessary with other stakeholders within Intel. Once we reach consensus within Intel, we try to influence the industry groups. The CFWG has imposed the requirement on the Intel group that their solution must ultimately be the industry standard, so there is give and take until we're all in agreement. It is the case today that for every topic of interest to us there is a parallel group at I300I, and the I300I group then takes the topic to their counterpart in Japan, the J300. In parallel, if there are SEMI standards to be voted on, a few individuals at Intel will begin an exhaustive campaign to ensure high SEMI voter turnout. This is important because a low turnout invalidates the vote, slowing down the

[7] www.sematech.org/public/division/300/metrics.htm

[8] Early on, manufacturers and suppliers alike had expressed the view that the SEMI process was too slow to be effective for defining standards at the beginning of the transition, rather than after the fact as had happened before. To date, the SEMI process has worked well enough, and all parties have accepted the fact that I300I requirements may precede formal SEMI acceptance, but that the goal remains to obtain SEMI standards status.

[9] www.sematech.org/public/division/300/guide.htm

setting of standards. Everything is in order and successfully exercised for the standards defining AMHS interfaces, lot carriers, mini-environments, some assembly issues, and others. More are still being worked on.

New and Improved Business Processes

Within Intel, a number of business processes have been improved or invented to ensure that the overall goals for the 300mm transition are met. A few deserve special mention, although most details are beyond the scope of this paper.

Intel has a time-tested equipment-selection methodology that is managed by the strategic committees mentioned earlier. Because at 300mm there were a number of new requirements that do not pertain to 200mm equipment, we developed a new set of selection-training materials and delivered them to key people from every selection team. The new selection methodology and requirements augment the existing system; they do not replace it.

Intel uses a purchase spec to define the requirements of our equipment and to manage supplier performance to those goals. The generic form of this document was updated to be consistent with the requirements that eventually became the Global Joint Guidance. Some other changes were included too, where those changes would drive increased productivity of the toolset. For instance, we changed the specification for individual tool scrap rates to be consistent with results already achieved within Intel, thereby increasing the overall line yield targets by approximately 2%. The achieved results had not yet been, and likely never would be, incorporated into the 200mm purchase spec. The writing of the new spec served as yet another scrubbing of our requirements, which ensured that we asked for industry standards, not some unique ones favored by a special interest group within Intel.

Several sources of data are used to set targets for tool performance, to provide real time measurements of tool performance, and to order tools based on some combination of the above. The sources tend to be in conflict, and there is no formal system to resolve the conflict, sullying the decision making process. For 300mm, we put in place a single database called the Selection Database (SDB), which serves as the sole source for decision-making data. The stakeholders have defined and agreed on a process for updating the database and for using it for various applications. The system has some shortcomings, e.g., it is not particularly user friendly, and we find ourselves needing to remind people that this is the sole source. But, it does work, and it enables engineers and managers to quickly answer questions and develop summaries that would otherwise be obtainable only by extensive scurrying.

Each selection team has the responsibility to develop cost models to manage suppliers to productivity expectations and AMHS interface requirements, to measure utilities and natural resource consumption, and to do several other tasks. While we had trained the teams on the requirements, we needed to oversee them on their performance to a degree never done before. If we failed to do that, we could be pretty sure that the full benefits of 300mm would not be realized. We put in place the Selection Synergy Working Group, chartered by the CFWG, whose initial function was to provide a single forum for answering questions about requirements and 300mm business processes. Somewhat later its function changed to ensuring that the teams were functioning properly, that they had membership from all the right groups at Intel, and that they had schedules for testing. Finally, today, the Synergy WG's role is to monitor progress on key deliverables of the selection team, to delve as deeply as possible into technical issues related to meeting 300mm targets, to challenge the teams to exceed the targets, and to provide clear summaries to management of performance across the toolset. For instance, at 200mm, it would be an exhausting effort to put together data on a single topic, such as, compliance to emissions' standards, across all the tools being selected within a given timeframe, even though in 200mm only a few tools are being selected at once. Through the action of the Synergy WG, we get summaries across the entire toolset on each of approximately ten different topics. In principle, the functions of the Synergy WG could be completed by the committees managing selection, but they have asked the CFWG to manage these details for them, while retaining responsibility for making the selection itself.

For 200mm selections, each team is individually responsible for securing the test wafers it needs to properly exercise tools. The test wafer flows are typically run in our development fabs. Since Intel has no 300mm tools, running test wafer flows is impossible so we replaced the distributed model of test wafer acquisition with a centralized model called the Silicon Clearing House (SiCH). The SiCH had three primary functions: (1) to determine the flows needed for all selections, (2) to secure processing of these flows outside Intel, and (3) to manage the logistics of pushing wafers from shipping and receiving to sites where processing could occur (primarily Austin, Santa Clara, and Japan), and then redistributing the wafers to the teams as needed. In this fashion, approximately 5000 wafer passes have been processed to date (Figure 6), with another 5000 anticipated prior to completion of all the selections. This is the most complex and underappreciated task of the entire 300mm program.

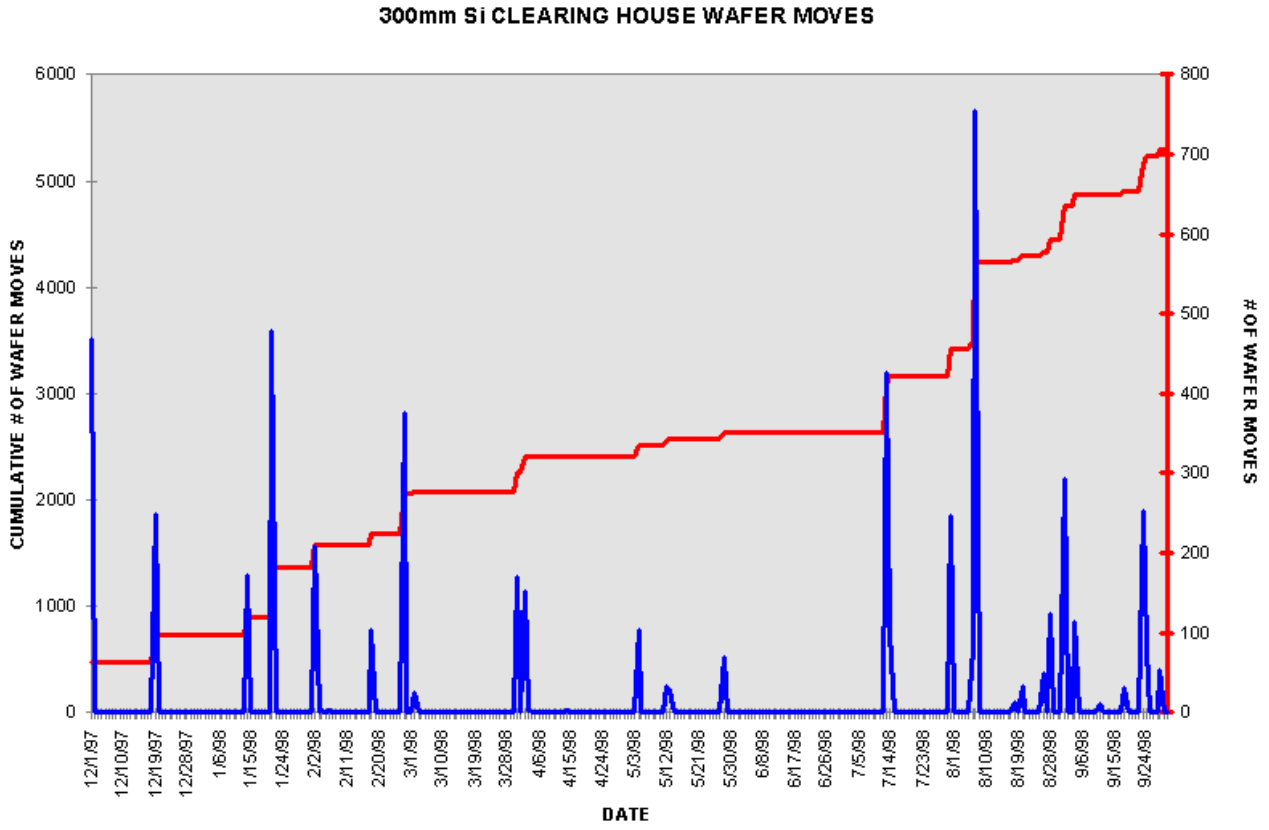

Figure 6: 300mm silicon clearing house wafer moves. Figure courtesy of Melton Bost.

The strategic committees managing selections are limited to examining the most expensive tools. The number of tools they manage is counted in tens, whereas the number of line items in the full catalog of everything needed for a factory is counted in thousands. Intel's change control policy, called Copy EXACTLY!, demands that every item be the same in all factories running the same process technology. The business processes for ensuring compliance on the less expensive tools were *ad hoc* at best; many decisions were not reviewed and were left up to individuals. For 300mm, the CFWG was asked to manage the selection of the thousands of tools falling in the cracks between the strategic committees. Many of these items are not wafer-size dependent, so they default to today's choices. Many others fail to meet criteria related to process sensitivity or total cost, so teams are asked to follow a formal selection process, but the results are not reviewed. The remainder, again numbering in the tens, will be reviewed by the CFWG as we near selection.

Getting to GO!

While the CFWG and the various other committees staffed by middle management served to set most direction and manage most details, for Intel to make the multi-billion dollar decision to go to 300mm, a higher level committee was needed. In July 1995, the first 300mm Steering Committee was formed, with a charter to make a recommendation for action based on an analysis of the full set of relevant issues. Comprised of vice presidents and senior managers representing engineering, materials, development, strategic planning, manufacturing, and finance, we made our first recommendation in early 1996: Intel should proceed towards 300mm, intercepting it at our 0.18um generation in Santa Clara, to be followed quickly by the 0.13um generation in Portland.

A second similarly stacked committee was formed to steer the project into existence. A date for first deliveries was set, October 1, 1998, and planning commenced. The new steering committee defined a process of reviewing overall status (or risk), and the First Risk Assessment was done in October 1997. We determined that Intel's need for additional capacity and the industry's ability to deliver it at 300mm no longer coincided. The committee pushed out our intercept to 0.13um, with first tools scheduled for an April 1, 1999 delivery. At the time, it appeared that more than six companies would put together pilot lines in 1998 and 1999.

Because the intercept no longer called for development at two generations and two sites, the committee task was greatly simplified, so it made itself once again. The Second Risk

Assessment was done in May 1998 and showed that the international slowdown signaled by the Asian financial crisis of 1997 had decimated the industry's plans for 300mm in 1998 and 1999. Only Siemens' pilot line, with partnership from Motorola and money from the German government, was likely to get off the ground in that period. Nevertheless, overall tool readiness appeared somewhat better, but real knowledge of it was too thin to make a startup decision. Furthermore, the schedule for the 0.13um intercept did not really demand first tools in April 1999; we could slip the first dock dates without compromising the process certification dates. Combining this with the slowdown in business and the reexamination of all major capital programs, the recommendations of the Second Risk Assessment group were (1) that we make no commitment to a startup date, (2) that we take the next six months to make detailed inquiries into tool readiness, and (3) that we do a Third Risk Assessment towards the end of 1998 to determine what date, if any, would work.

Fourth Quarter of 1998 Outlook

The overall industry slowdown has reached historic proportions, and the lack of a startup date has made management of the suppliers more difficult than usual, but not impossible. We have weathered the dog days of summer 1998 when layoffs in the equipment industry became routine, when the world's largest supplier withdrew from I300I undermining I300I's importance and seeming to threaten it with extinction, when at SEMICON West senior managers from various sides met but seemed near fisticuffs, when banner headlines in the San Francisco Bay area papers announced the death of 300mm, when every pundit, no matter how small, offered their opinion on the demerits of the industry transition, and when our own selection teams declined to fulfill program commitments because they felt management was itself not committed to continuing the program. Ignoring all the negativity, we are focussing on the critical task of getting the data on tool production worthiness, by working extensively with I300I, with suppliers, and with key groups at Intel.

As we enter the last quarter of 1998, we are vigorously exercising a process to review the teams' equipment data tool-by-tool, anticipating the Third Risk Assessment before year's end. As of this writing, we expect that approximately 80% of the required tools will be ready for selection and a January 2000 startup. As for the remaining 20% of the tools, there are no known showstoppers; they will just need to be managed carefully.

Overcapacity is driving manufacturers to extend the useful life of 200mm equipment where ever possible. The 200mm era, in the sense of new construction of 200mm lines by the major manufacturers, is effectively over. The next big opportunity for the suppliers is 300mm. The forecasted productivity of the 300mm equipment looks very good, appearing to come close to capital productivity targets, as shown with historical perspective in Figure 7. If the forecasts are borne out in practice, then 300mm will seem a much more attractive alternative to using 200mm fabs, even those where the equipment set is largely depreciated. When the realization of this counterintuitive result sinks in, as it has here, the industry will make the transition much more rapidly than it did when it went to 200mm[10]. The first year in which this will have a big impact on suppliers' revenue is 2000, and the first year it will have an impact on manufacturers' cost is 2002, possibly as late as 2003. This impact, as measured by die cost reduction, will be twice as large as that of any wafer size conversion in our short history. The two lessons we should take away from this are that such transitions should be planned as an industry and that wafer size should increase by at least 50%.

[10] The transition to 200mm took place over many years. IBM began production in 1988, but the transition was still gaining momentum in 1993 when Intel began 200mm production.

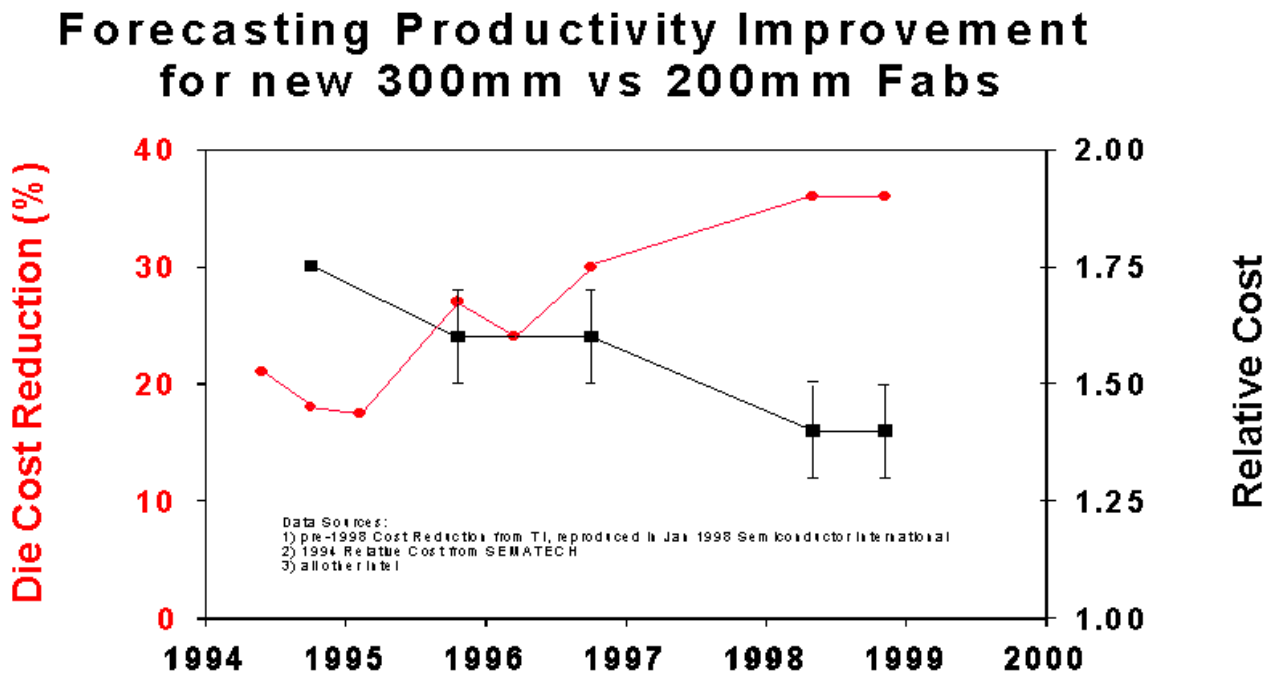

Figure 7: Forecasting productivity improvements for new 300mm vs. 200mm fabs

Acknowledgments

First, I would like to thank the early members of the Cross Functional Working Group, particularly those who articulated the automated materials handling roadmaps and gave us all a model to follow. These individuals include Sandra Viarengo, Dev Pillai, Srini Srinivasan, Bala Subramaniam, Jeff Pettinato, Fred Voltmer, Richard Parker, Noel Acker, and Yolanda Leung. Next, I would like to thank my other colleagues who have over time helped build the infrastructure that is now in place and working today. These include Winston Saunders, Carolin Seward, Dave Krick, Melton Bost, Larry Norman, Tom Garrett, Ken Moyle, Gordon McMillan, Rich Poliak, William Yeh, Roy deGroot, K.V. Ravi, Diana Harris, Brian Hunter, Stefan Radloff, Stephen Sumner, John Morrissey, Derek Youngson, Dan Enloe, Tom Abell, John Souza, and Kwang-Woo Choi. Also, Ann Marie Kenitzer and Peter Silverman have supported the supplier management and tool selection effort, providing staffing and invaluable critique throughout. Frank Robertson, the General Manager of I300I deserves special mention. There are many others, and I apologize to those I have left out unintentionally. Finally, I would like to acknowledge the contributions of Intel senior managers who provided inspiration, time, support, guidance, and patience over what has now become a long development effort. These include Gulsher Grewal, Don Turner, Don Rose, Paolo Gargini, Ken Thompson, Bob Jecmen, and Youssef El-Mansy.

Author's Biography

Dan Seligson spent his first ten years at Intel as a researcher in Bubble Memories, X-Ray Lithography, and Neural Networks. During this period, he did a two-year assignment at SEMATECH in Austin and was a researcher in residence at the Hebrew University in Jerusalem for six months in 1990. Subsequently, in 1994 he began managing equipment suppliers in the Thermal Processing Area and building the 300mm program. Today he is the Front End Synchronization and 300mm Manager in TME. Prior to joining Intel, he received a B.Sc. from MIT in 1977 and a Ph.D. from Berkeley in 1983, both in physics. He holds five patents and has a long publication list. Outside family and work, his interests include windsurfing, music, and photography. His e-mail is daniel.seligson@intel.com.

Improving Throughput Across the Factory Life-Cycle

Karl G. Kempf, TMG/TME Decision Support Technology, Intel Corp.

Index words: constraints, capacity, supply

Abstract

A semiconductor factory goes through many phases in its life cycle including design, build, various ramps, and many levels of production. Maximizing the profitability and return on investment across this life-cycle is a critical component of Intel's approach to financial success. We have been applying the concepts of Goldratt's Theory of Constraints across the factory life-cycle and have realized improved performance in many of these phases, as well as in the integration of the phases.

The Problem

Within Intel Corporation, there are at least three identifiable supply lines (Figure 1). The most obvious from outside the company is the product supply line. This supply line includes planning to schedule production, materials to supply the ingredients, manufacturing to produce the products, and logistics to deliver them. This is the supply line that springs into action when you place an order with Intel and find it being delivered a short time later. Another supply line for which Intel is famous is its technology supply line, which has two major branches. One is product design, delivering a stream of ever faster and more capable product designs for manufacturing to build. The other is process design, providing a sequence of ever finer and more capable processes for manufacturing to follow in building products. Together they form the supply line that responds to the insatiable market demand for faster semiconductor devices with higher functionality. Perhaps the least obvious supply line from outside the company is the capacity supply line. This is the supply line that manages manufacturing resources. It is always trying to supply the most cost-effective manufacturing capability synchronized with market demand. This supply line involves at least the selection and layout of equipment (design), construction of buildings (build), startup of production (ramp), and operation (Mfg) of Intel factories. Since managing this capacity supply line is the primary focus of this paper, we discuss how this supply line is driven, how its components work together, and what problems it must overcome.

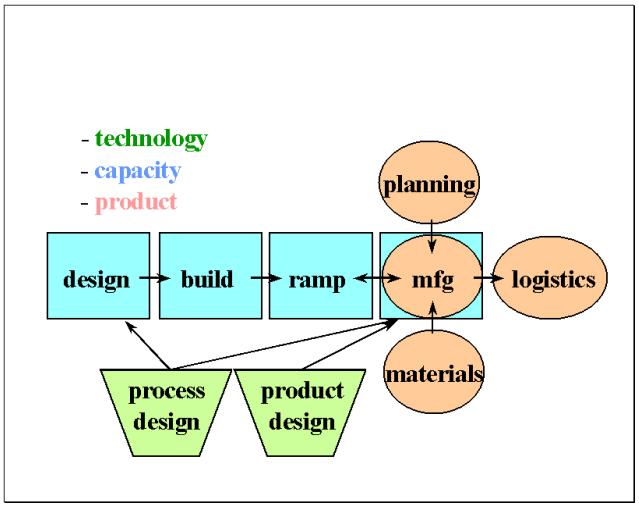

Figure 1: Intel supply lines

Clearly the capacity supply line is driven by market demand. More interestingly, it is also driven by advances in process technology. As semiconductor process design moves to finer line widths, new cleaner factories with improved equipment are needed to produce cutting-edge products commanding premium prices. This results in older factories being relegated to run products that are no longer at the cutting edge and which the market treats as commodities. Both these scenarios have a profound effect on the capacity supply line. Because the expense of building new factories is rising rapidly, there is great pressure to keep costs as low as possible and get productivity as high as possible. Because the older factories are now producing commodity products where every penny counts, there is also great pressure on them to keep costs as low as possible and get productivity as high as possible. One of the major themes of this paper, therefore, is "doing more with less" in the capacity supply line.

Another major issue for the capacity supply line is integration. It is surprising to note that the four main tasks in the design-build-ramp-run sequence of the capacity supply line can be executed almost independently with very little information flowing between them. While blindly following this factory life cycle sequence will result in a running manufacturing facility, if the tasks are not well integrated, the facility

might require extra time and/or money to complete. And while the resulting manufacturing facility will produce products, it might not do so very efficiently from a time or cost perspective. Integrating the design-build-ramp-run tasks provides a dual benefit for the capacity supply line: supplying the capacity as efficiently as possible and applying that capacity to supply products as efficiently as possible. Therefore, another major theme in this paper is "efficiency through integration" in the capacity supply line.

Finally, almost every step in the design-build-ramp-run sequence that makes up the capacity supply line involves the variable availability of resources. Most of the activities in the build, ramp, and run tasks require the simultaneous availability of equipment, materials, and skilled personnel to progress. The absence of any one resource stops activity, and the availability of all such resources is variable. Equipment breaks and needs to be maintained. Materials are supplied by vendors with imperfect resources. People take breaks and even when working diligently can only be in one place at a time. Even the design task requires detailed data about the variability in the availability of the build, ramp, and run resources. The third major theme of this paper, therefore, is "managing variability."

The problem for the capacity supply line is to supply the most cost-effective manufacturing capability synchronized with market demand. Three themes interact to complicate any approach to managing the tasks involved in solving this problem. The pressure to do more with less is never ending and takes different forms over time. The individual tasks are complex enough that it is tempting to try to divide and conquer them, but not tackling them as an integrated whole will prove very expensive in time and money. The availability of resources for each of the tasks is always variable.

The Basic Solution

Over the past several years, we have been able to employ the concepts of Goldratt's Theory of Constraints (ToC) to improve the performance of our capacity supply line. The most abstract version of Goldratt's ToC has to do with making money. The most concrete version has to do with managing individual resources. Both versions are summarized here and then applied to tasks in the capacity supply line.

One of Intel's corporate goals, supported in different ways by each of the supply lines, is to make more money now and in the future. Moneymaking is usually measured with two parameters: net profit (how much did we make) and return on investment (relatively, how much did it cost us). Ideally Intel maximizes profit while minimizing the investment required.

Translating these ideas into capacity supply line terms, we can use this corporate goal to drive supply-line decision making. Throughput (T) is money generated by manufacturing that is directly related to quality product shipped on time resulting in sales. Some expenditures are required to make T. Inventory (I) is money inside the capacity supply line such as equipment and spares and in-process materials. Operating Expense (OE) is money required by the capacity supply line such as overhead and personnel expenses to turn inventory into sales. These terms are related to profit (P) and return on investment (RoI) as:

$$P = T - OE$$

$$RoI = (T - OE) / I$$

These equations explain the pressure to reduce inventory and operating expenses while increasing throughput in all stages of the capacity supply line as an approach to doing more with less.

ToC derives its name from the key observation that in any system, the resource with the lowest capacity constrains throughput. The key process in ToC is aimed at improving throughput as the best way of driving up profit and return on investment. Step 1 involves identifying the system constraint. Step 2 focuses on understanding all means to exploit the constraint and maximize its throughput. This almost certainly includes protecting the constraint from the variability of other resources. Step 3 subordinates all other resources to the constraint, supporting all means of exploiting its capacity. Step 4 advises that whenever possible the constraint should be broken or removed, raising the throughput of the system, and the improvement process rejoined at Step 1.

The rest of this paper describes how we have used these simple ideas to develop powerful techniques to integrate and optimize the Intel capacity supply line.

Manufacturing

Our first and so far most successful application of ToC to the capacity supply line has been in manufacturing. This was an obvious place to start since manufacturing is included in all of the supply lines shown in Figure 1. Consider the simple factory shown in Figure 2. There are three processing steps, each with a machine, an operator, and an average run rate in units per shift. Since Step 2 has the lowest capacity of all of the resources in the system, it is identified as the factory throughput constraint. The factory cannot produce any more than this step can run, and any time this step is idle, factory capacity is irreversibly lost. As part of the exploitation process, its rate is identified as the "drumbeat" with which to synchronize the rest of the production line. To fully exploit

the capacity of the constraint, it must have three things available at all times: material to work on or work-in-progress (WIP), a machine to load the product into, and a skilled operator to perform the work. Subordination of the rest of the resources of the factory involves ensuring the constraint has its requirements satisfied at all times. If the factory capacity is to be raised, the capacity at Step 2 must be raised by improvement projects or equipment acquisition. And if it is raised beyond 900 units per shift (ups), then Step 2 is broken as the constraint and Step 3 takes its place.

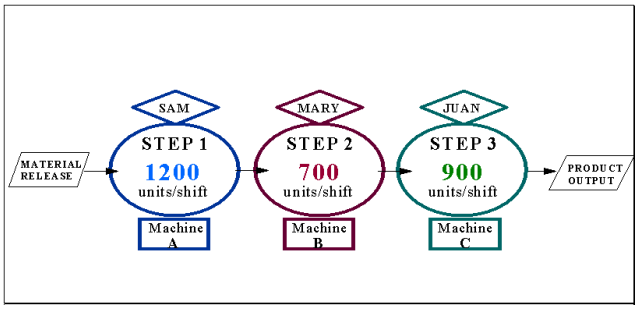

Figure 2: A simple factory

The first requirement for exploitation of the constraint is WIP, and other resources must be subordinated to ensure that the constraint is always fed. One cause of the constraint starving is the inevitable breakdown of Machine A (variable availability). One way to protect the constraint is to place a WIP buffer between Machine A and Step 2. The size of the buffer is based on the historical distribution of times to repair Machine A. Machine A is subordinated to Step 2 by always being run in such a way as to maintain the correct level in the buffer. Too much WIP in the buffer raises factory throughput time (TPT) but does not raise output. Too little WIP in the buffer risks factory capacity. Any other way of running Machine A fails to optimize constraint performance and therefore factory performance.

Another way to ensure that WIP is fed to the constraint is to control material release. Naively releasing 700 ups is not adequate. Subordinating material release to the constraint involves allowing the constraint to pull in the amount of work it requires. When the constraint is undergoing maintenance, less material is released. When the constraint exceeds its average output, more material is released. This concept is described in ToC as tying the "rope" between the constraint and material release so that the constraint can "pull in" the work it needs as it needs it.

The ToC-based approach to WIP management gets its name from the combination of these ideas: drum-buffer-rope or DBR (Figure 3). Consistent use of these ideas drives the factory towards maximum throughput at minimum throughput time in the face of any variability in the availability of equipment. We have used these basic ideas, complemented with our own extensions, in process technology development facilities (TD fabs), high-volume manufacturing fabrication facilities (HVM fabs), and factories where die are assembled with their packaging and final testing is done (A/T). In all cases, factory throughput has gone up by 10% to 20% while inventory has gone down, with no capital outlay or increase in operating expenses. Simply getting the same equipment and people to work more effectively is the key.

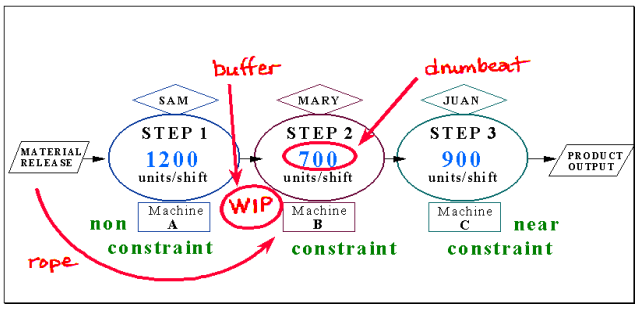

Figure 3: The Drum-Buffer-Rope factory

The second requirement for exploitation of the constraint is that the constraining equipment be up as much of the time as possible. The subordination required here is that maintenance of the constraint equipment takes priority over other equipment. In general, the more the equipment constrains the factory, the higher its maintenance priority. In the simple factory, Machine B has priority over Machine C, and Machine C has priority over Machine A. Since Machine B being down is equivalent to the factory being down, we have extended this basic thinking to further specify the number and training level of personnel dispatched to perform maintenance as a function of the constrained-ness of the tool(s) involved. This approach can yield as much as 5% more throughput with no change in inventory or operating expense.

The third requirement for constraint exploitation is that the constraint be staffed at all times. The subordination here can come in many forms. For example, machine operators assigned to the constraint must collaborate to cover for each other at breaks. Furthermore, breaks for machine operators should be coordinated with the activities of repair technicians. Another possibility is to cross-train operations and maintenance personnel on multiple tasks. The subordination here is to bias cross-training more toward the constraint and less toward

the non-constraints. In the simple factory, all three operators would be trained for Step 2, two operators for Step 3, and one for Step 1.

These ideas can be applied to many other facets of manufacturing to further manage variability and get more throughput from the same (or less) inventory and operating budget. For example, consider line yield (the scraping of WIP at intermediate positions in the process flow). With finite resources to address line yield problems in the simple factory, it is important to address line yield losses at Step 3 before Step 1. This is true because losses at Step 1 can be recovered by releasing more raw material into the factory and using the excess capacity at Step 1. Losses at Step 3 are much more costly since each involves irreversibly discarding the constraint capacity invested in the WIP that is lost, and running new material through the constraint again to replace it.

Another example involves the prioritization of engineering projects (assuming finite engineering resources). Given the choice of increasing the capacity of Machine A through an elegant and interesting engineering effort, or decreasing the preventive maintenance time on Machine B through a mundane and uninteresting effort, the latter should have priority. The reason for this is that the mundane project will increase factory capacity by increasing the availability of the constraint to do productive work. Investing effort in the former project on a non-constraining tool will have little (if any) positive impact on factory performance.

Design

Our second major application of ToC to the capacity supply line has been in the area of factory design (Figure 1), specifically in selecting the number of pieces of equipment to purchase. The goal of "doing more with less" in the capacity supply line starts here. The naive approach to design would simply be to build a balanced line, that is, a line that has the same capacity at each process step and is equal to the desired output of the factory (Figure 4). Aside from being very difficult because of the integer nature of equipment, ToC argues that this would be a very difficult factory to operate. Although there is no obvious constraint from the point of view of inspection of the average run rates of the three processing steps, there would be a constraint on the floor of such a factory. It would be the last machine that had an unscheduled breakdown, and so would move around. This would make it very difficult to instruct Sam, Mary, and Juan on how to run shift by shift, to release the right volume of material, to optimize cross-training, and to prioritize equipment maintenance, line yield improvement, and engineering projects.

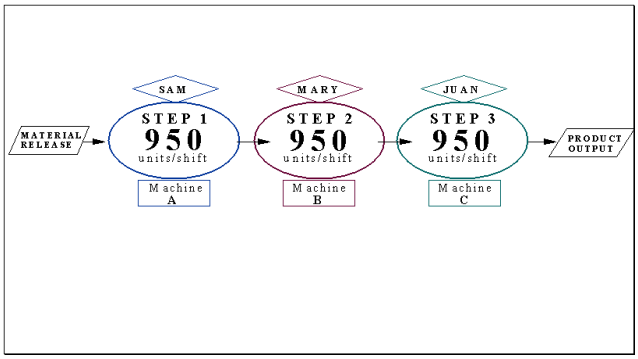

Figure 4: A balanced line

ToC would argue for an unbalanced line. Such a line might require a small increase in equipment inventory to provide the imbalance that identifies the constraint, but would supply increased throughput due to its ability to be efficiently managed. The interesting question that we have answered over the past few years is "how much imbalance?" Too little imbalance in the design of factory capacity leads to inefficient operation and lost throughput. Too much imbalance leads to wasted capital expenditure and too much equipment inventory, not to mention too much of an operating expense to run and maintain it.

The first key to answering the imbalance question has to do with the cost of the equipment. Reasoning from the inventory perspective, the most expensive equipment set should be the constraint, the next most expensive equipment set the near constraint, and so on. This means that we can imbalance the factory with the least expensive equipment sets making them the non-constraints.

The second key to the imbalance question has to do with the manner in which the variability in the availability of the equipment, people, and WIP stacks up across the factory. Since it is important to protect the constraint from starving due to the variability of upstream resources, more of those highly variable resources can be purchased. This imbalances the line, and does so in such a way as to reduce the variability of the troublesome equipment set by having more members in the set.

The third key to the imbalance question has to do with the WIP management ideas described in the previous section. Remembering that the goal of drum-buffer-rope is to maximize throughput at minimum throughput time in the face of any variability in the availability of the resources, it is important to include the WIP policy in the factory design process. Selection of the equipment sets to be the constraint and near-constraints must include considerations of how easily they can be exploited and how easily the other equipment sets can

be subordinated. As a contingency for future success and market upside, the design process must also consider the possibilities for breaking the constraint.

After practicing on two previous process technologies, we have applied this approach to our three latest process technologies as they rolled out and forced existing factories to be refit with new equipment or else new factories had to be built. The savings compared to our previous design methods have been in the range of 3% to 8% in capital cost (or I) for equal or improved throughput. While these percentages may seem small, given the multi-hundreds of millions or billions of dollars spent on equipment in each of our factories, the absolute savings have been substantial.

Ramp

Our most recent application of ToC to the capacity supply line has been in the area of factory ramp (Figure 1). Once again, the theme of "doing more with less" arises, and once again ToC points a way forward. Ramping a factory means going from one level of production to another. In an existing factory, this might mean ramping volume up or down on the current process, or ramping down an old process and ramping up a new process. In a green field situation, this means going from zero production to full-volume production. Since semiconductor production equipment is so expensive, it is normal in all these cases for equipment installation and equipment operation to be going on simultaneously. Once one (or a few) pieces of each type of equipment has been installed, raw materials are released into the line and production proceeds. As more tools are added, more raw materials are released.

ToC can be used in two ways to produce the fastest, cheapest ramp. One is to determine the identity and utilization of the constraint. On the one hand, since production is going on during the ramp, all of the ideas described above in the section on manufacturing should be applied for maximizing throughput. This would require that the constraint is known and does not move around. On the other hand, equipment is being installed daily and the capacity of the factory is dynamic. This means that the identity of the constraint could change from day to day. Goldratt's ToC argues that the installation of the equipment should be choreographed so that the identity of the constraining equipment set is constant and throughput can be maximized.

The other way in which ToC can be applied is to use the same thinking described in the section on manufacturing, but this time, transform the entities being discussed. In manufacturing, the process flow is the physical/chemical transformations being made on the wafers in fab or the die in A/T. The WIP is the product that is moving across the flow. The constraints are usually the processing equipment or the personnel operating or maintaining the processing equipment. During a ramp, the process flow is the installation and qualification steps required for each of the equipment types. The WIP is the pieces of processing equipment being installed including supporting materials. The constraints are the electricians, piping specialists, mechanical contractors, qualification technicians, and so on who are executing the installation and qualification steps.

Once this transformation has been made, it is simply a matter of using the identify-exploit-subordinate-break process that is the core of ToC. The constraining resource is identified, and all ways to exploit it for maximum throughput are identified. All other resources are subordinated to the constraint. If higher throughput is desired, that is, the ramp needs to be done faster, the constraint must be broken.

This leads to the interesting question that we are currently answering, that is, "How fast should a ramp be done?" At one extreme, a large number of resources could be utilized (high OE), and the ramp could move along very quickly. But, if the result is more product at a faster rate than the marketplace can absorb (low throughput), then the RoI on the ramp is not very good. (Remember T is product sold, not just product produced.) At the other extreme, a very low-speed ramp could be executed with few resources (low OE). But, if the result is less product at a slower rate than the marketplace can absorb (low T), again the RoI is not very good. This is magnified by the fact that inventory would also go up since the equipment would be WIP for a longer period.

And of course, the ever-present variability in the availability of the resources has an impact on the ramp rate just as it did on the design and manufacturing phases. However, we expect that applying Goldratt's ToC principles to this balancing problem in the face of variability will increase our performance during a ramp by as much as 15%.

Integration

Each of the previous sections has described how we have used ToC to do more with less while managing variability across the capacity supply line. The topic that has not been mentioned since the problem statement is "integration," and ToC has helped in many ways on this important topic.

Using ToC in the design phase has decreased the amount of equipment we purchase (I) to deliver the same throughput (T). This means that even if we didn't use ToC in the ramp phase, there would be less equipment to install and so the ramp could be faster and cheaper. But since we do use ToC to ramp, we can better manage the constraint and apply all of our manufacturing ToC ideas earlier to realize higher throughput sooner from the reduced equipment set.

Using ToC in the design phase forces one to include the WIP policy that the resulting factory will use, and that in turn forces one to carefully consider how the factory will run at high volume. The fact that the WIP policy has already been designed means that it can also be applied during the ramp instead of waiting until all of the equipment has been installed.

Last but not least, we use modeling and simulation to try different scenarios around the ideas of ToC over the factory life cycle. And in some cases, we embed ToC ideas into our automation systems. The fact that one model can be implemented and used during the design phase, and the same model can be used in the ramp phase, and used again in the manufacturing phase saves a tremendous amount of effort and provides a very large boost to continuity. For example, we are now in a position based on our work with ToC to design a WIP policy during the design phase, and to plug it into our automation system to be used in the ramp and manufacturing phases. This is a markedly different approach from that of considering the factory life cycle phases as separate minimally-communicating events.

Conclusions

The application of ToC to the design, ramp, and manufacturing components of Intel's capacity supply line has significantly benefited each individual component financially as well as benefiting the integration of the components. ToC has been a practical way to continuously improve the "more for less" mentality that pervades our capacity supply line, and it has enabled us to manage the inherent variability of availability of all of the resources that the capacity supply line depends upon.

The most obvious missing component of this story is the application of ToC principles to the build component of the capacity supply line. Thinking even more broadly, one might speculate about the magnitude of the benefits of applying ToC to the non-manufacturing components of the product supply line and the technology supply line. Pushing outside of the supply lines, one might inquire about using ToC in finance, human resources, or marketing. Given the steady increase in our rate of applying ToC based on our successes, it should not be too long before a description of our work in this area appears in this journal.

Acknowledgments

There are far too many contributors to this work to list here. The intellectual leaders are Bruce Sohn, Eamonn Sinnott, Bob Rodgers, Steve Notman, Ray Dudonis, Ken Gray, Greg Mazenko, John Bean, and Dane Parker. Management support has come from Mike Splinter, Bob Baker, David Marsing, Chuck Roger, Don Rose, and Gene Meieran.

References

1) E. Goldratt and J. Cox, "The Goal," North River Press, 1984.

Author's Biography

Karl Kempf is currently the Principal Scientist for Manufacturing Systems for Intel Corporation in Chandler, Arizona. He holds degrees in physics, chemistry, and applied mathematics/computer science and is interested in performance optimization in complex man-machine systems. While working at SEFAC Ferrari, he was a member of three world championship teams. At Pinewood Movie Studios, he was on the team that won an Academy award for Special Cinematic Effects for the Superman series of movies. His e-mail is karl.g.kempf@intel.com.

Line Defect Control to Maximize Yields

Sanjiv Mittal, CTM/D2, Intel Corporation
Peter McNally, CTM/D2, Intel Corporation

Index words: defects, yield, cost

Abstract

This paper discusses line defect control through the use of defect monitors in semiconductor manufacturing. Defect monitor development has focused on maximizing good die output through die yield improvement in a cost-efficient manner. Line defect monitors provide rapid feedback and shorten cycle times for problem resolution. For high-volume manufacturing, line defect control is employed to achieve rapid excursion response and more stable yields. Return on investment analysis optimizes the cost of defect metrology against die cost reduction achieved by higher die yields.

Introduction

A semiconductor factory must provide predictable output to meet its customer commitments. Predictable output is based on meeting the die yield, line yield, and wafer throughput time forecasts for the factory. Today's semiconductor manufacturing processes have over 150 process steps and several weeks of throughput times. Several weeks of output is at risk if the only metric to measure the quality/yields of the wafers is at final test. Line defect control is a method that uses inline defect monitors to measure defect/quality levels on product wafers at various sampling points throughout the manufacturing line [1,2]. Inline defect monitors give quantitative and qualitative information about the types of defects detected on the wafer surface. A response system based on the information collected by these defect monitors assures good line defect control.

Line Defect Control Method

Defect monitor development has focused on maximizing output through die yield improvement in a cost-efficient manner. Numerous manufacturing and die yield advantages from product wafer defect monitoring have been realized over traditional bare test wafer monitoring.

Line defect control is achieved by measuring and controlling defects on process equipment/tools and by inspecting and controlling defect levels on product wafers. Defect levels on process equipment are measured by running silicon test wafers through the tool and then responding to shifts in defect levels measured on the test wafers. The traditional method of inspecting product wafers was to visually inspect wafers under a microscope. The visual inspection technique worked well for detecting relatively large and high-density visible defects. However, it was limited by the skill of the inspector and the limited die sampling area. As process technologies move to smaller geometries, the size of yield-limiting defects is scaled with the feature size. Similarly, the high capital cost of current semiconductor factories requires ever-lower defect densities for each generation of technology. The industry has responded with the use of automated defect inspection equipment for product wafers and parallel development of response systems, called product line monitors.

Defect detection is a critical component of defect reduction and control (1). The SIA Crosscut Technology Working Group [3] has formed its own sub-group to ensure that the roadmap for future defect detection capabilities is consistent with the future (higher) yield requirements forecasted by the industry. The two primary techniques for automated defect inspection are either optical based or laser detection based. In the optical inspection method, pixels in one die or cell are compared to adjacent cells, with differences counted as defects. This method is very good for catching small visible defects and even subtle pattern variations. As the sensitivity is increased, however, optical detection is limited by noise from subtle color variations and natural thin-film grain structure variation. In the laser defect detection method, defects are detected from reflected signal changes as the laser is rastered over a surface anomaly. The laser method is good at picking up defects on layers where the wafers are relatively planar; that is, when the defects are large relative to the pattern topography. Laser tools scan much larger areas per unit time than do optical tools. Both methods have their strengths, and they are used accordingly to provide the best visibility of different types of defects. Once the die are inspected, optical reviews are used to classify defect types. Defect

paretos, as shown in Figure 1, are constructed to understand the different types and levels of defects at each inspection location.

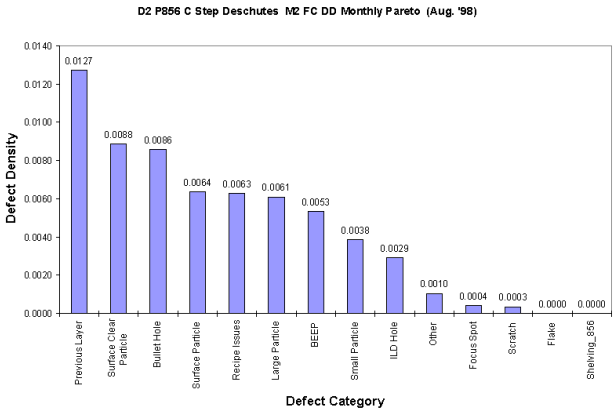

Figure 1: Defect classification pareto for a defect monitor

For line defect control, multiple product line monitors are inserted throughout the manufacturing process so there is continuous feedback on the stability of defect levels. Monitor locations are chosen based on multiple considerations including the excursion risks of preceding operations, the yield impact of the layer defect population, and the quality of the monitoring recipe.

Defect Control in Manufacturing

As with other commonly monitored parameters, statistical process control (SPC) is used to monitor defect trends and trigger responses, as shown in Figure 2. Automated factory floor response systems are initiated when defect limits violate statistically determined SPC control limits. These Out of Controls are the first line of defense against tool excursions.

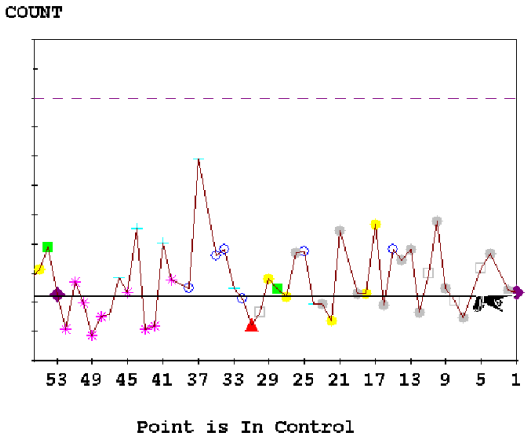

Figure 2: Defect monitor control chart

The response systems in a manufacturing line are critical to the effectiveness of the line defect monitors and to the workflow of the manufacturing line. If the manufacturing technicians (MT's) do not have a well defined response flow system, then either the response is inadequate or the workflow of the manufacturing line is interrupted while an engineer is called in to decide how to respond. The MT's must be trained to respond on a 24-hour basis.

Line defect control can be used to improve yields by adding stability to the line and by using the defect pareto data (Figure 1) to eliminate or reduce "killer" defects. Using SPC, line defect monitors can be used to detect defect excursions, and the rapid feedback minimizes material loss. Line monitors provide rapid feedback.

Cost/Cycle-Time Implications

Improved line defect control, and therefore yield control, can be achieved by increasing the number of defect monitors placed in the line. However, placement of defect monitors leads to additional cost and increases the cycle time of the process. The increase in manufacturing cost comes from the cost of capital (inspection tools) and labor in collecting and analyzing the data. The cycle time of the overall process increases due to the time taken to do the additional monitoring. For example, if a manufacturing cycle time is six weeks, and there are no line defect monitors in place, we could potentially have up to six weeks of material in jeopardy in a case where a die yield defect-related problem is detected at final test. However, if a monitor were placed in the middle of the manufacturing line, this would cut the amount of material in

jeopardy by half, assuming the line defect monitor could effectively catch the defect. Thus, the feedback loop for the line defect and yield control can be improved by strategically placing monitors throughout the line. Moreover, in an environment where we are willing to tolerate an output risk of up to one week of material, we should have five line defect monitors in the above six-week manufacturing cycle time example. Assuming a 0.5 day cycle time per monitor, placing five monitors would increase the manufacturing cycle time by 2.5 days and increase labor and capital cost. Figure 3 shows a conceptual cost relationship of product defect monitor frequency to die cost. The parabolic shape in Figure 3 is caused by high-defect metrology costs at short monitor distances on one end, and by low die yield, by potentially missing yield excursions due to large distances between monitors, on the other end. The lowest point in the curve is the optimum balance between die yield and monitor distance/frequency.

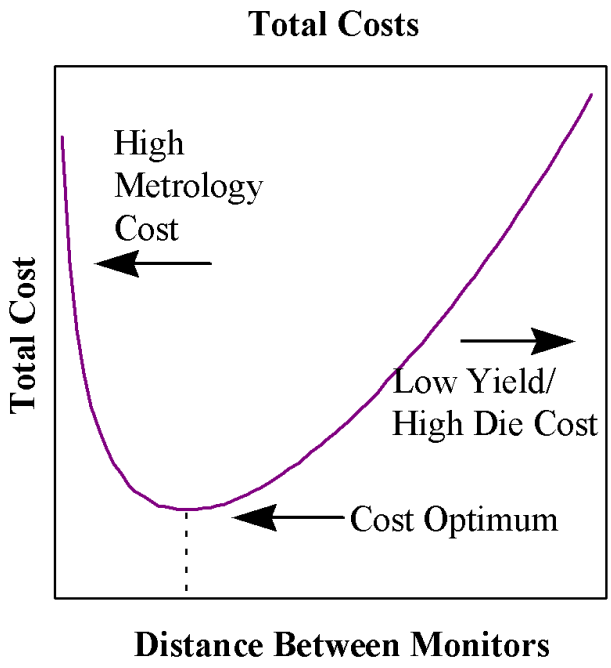

Figure 3: Cost relationship of the number and frequency of defect monitors in a manufacturing line

Future Trends

In the future, we hope to seamlessly integrate line defect inspection, data analysis, and response systems into the semiconductor manufacturing line. There are three significant trends for line defect control: more sophisticated automation to collect the defect description data, improved response to line defect control data, and more emphasis on cost reduction. Automatic defect classification will enable us to significantly reduce the labor involved in collecting and summarizing the current defect data. Improved automation systems and algorithms are being put in place to allow a faster automatic response to "defects of interest." For example, we envision a time when a defect source is known and there is an increase in defect levels from that source, there will be an automatic response signal sent to the suspect station or stations. The trend in cost reduction will force us to reduce the overall level of monitoring in a semiconductor process and make it more efficient.

Line defect control is now an essential part of the semiconductor manufacturing process to maximize die output. As we look forward, we should be able to improve the level of line control and reduce costs by increasing the sophistication of the tools, improving our response systems, and reducing the total number of monitoring points.

References

[1] C. Weber, D. Jensen, E. Dan Hirlemann, "What Drives Defect Technology," *Micro*, June, 1998, pp. 51-75.

[2] S. Mittal, K. Lubic, P. McNally, "Use of Inline Defect Monitors to Drive P852 Yield Improvement," *Intel Manufacturing Excellence Conference*, 1995.

[3] *The National Technology Roadmap for Semiconductors*, San Jose, SIA, pp163-178, 1997.

Authors' Biographies

Sanjiv Mittal has been at Intel since 1984. He is currently the Fab Manager at Intel's D2 Development and Manufacturing Fab in Santa Clara. Prior to this position, he managed yield improvement and then the manufacturing departments. He has a Sc.D. from MIT and a M.S. from Purdue University both in materials science and engineering and a B.S. from the Indian Institute of Technology. His e-mail is sanjiv.mittal@intel.com.

Peter McNally has been with Intel's D2 development and manufacturing facility since 1993. He has served various yield improvement roles driving defect reduction, metrology systems, correlation tools, and yield analysis for Intel's .5 and .25 micron processor generations. His 15 previous years were spent at Hewlett Packard and National Semiconductor including an assignment with Sematech. His degrees include a M.S. from Stanford University and a B.S. from Cornell both in materials science and engineering. His e-mail is peter.mcnally@intel.com.